IDENTIFICATION AND INFERENCE OF NONLINEAR MODELS USING TWO SAMPLES WITH ARBITRARY MEASUREMENT ERRORS

By

Xiaohong Chen and Yingyao Hu

November 2006

COWLES FOUNDATION DISCUSSION PAPER NO. 1590

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Identification and Inference of Nonlinear Models Using Two Samples With Arbitrary Measurement Errors*

Xiaohong Chen†
New York University

Yingyao Hu‡
University of Texas at Austin

First version: February 2006. This version: October 2006

Abstract

This paper considers identification and inference of a general latent nonlinear model using two samples, where a covariate contains arbitrary measurement errors in both samples, and neither sample contains an accurate measurement of the corresponding true variable. The primary sample consists of some dependent variables, some error-free covariates and an error-ridden covariate, where the measurement error has unknown distribution and could be arbitrarily correlated with the latent true values. The auxiliary sample consists of another noisy measurement of the mismeasured covariate and some error-free covariates. We first show that a general latent nonlinear model is nonparametrically identified using the two samples when both could have nonclassical errors, with no requirement of instrumental variables nor independence between the two samples. When the two samples are independent and the latent nonlinear model is parameterized, we propose sieve quasi maximum likelihood estimation (MLE) for the parameter of interest, and establish its root-n consistency and asymptotic normality under possible misspecification, and its semiparametric efficiency under correct specification. We also provide a sieve likelihood ratio model selection test to compare two possibly misspecified parametric latent models. A small Monte Carlo simulation and an empirical example are presented.

JEL classification: C01, C14.

Keywords: *Data combination, nonlinear errors-in-variables model, nonclassical measurement error, nonparametric identification, misspecified parametric latent model, sieve likelihood estimation and inference.*

*The authors would like to thank R. Blundell, R. Carroll, B. Fitzenberger, J. Heckman, S. Hoderlein, A. Lewbel, E. Mammen, L. Nesheim, S. Schennach, C. Taber and conference participants at the 2006 North American Summer Meeting of the Econometric Society for valuable suggestions. Chen acknowledges partial support from the National Science Foundation.

†Department of Economics, New York University, 269 Mercer Street, New York, NY 10003, tel: 212-998-8970, email: xiaohong.chen@nyu.edu.

‡Department of Economics, University of Texas at Austin, 1 University Station C3100, Austin, TX 78712, tel: 512-475-8556, email: hu@eco.utexas.edu.

1 Introduction

Measurement error problems are frequently encountered by researchers conducting empirical work in economics and other fields in social and natural sciences. A measurement error is called *classical* if it is independent of the latent true values, otherwise, is called *nonclassical*. In econometrics and statistics there have been many studies on identification and estimation of linear, nonlinear and even nonparametric models with classical measurement errors. See e.g. Fricsh (1934), Amemiya (1985), Hsiao (1989), Chesher (1991), Hausman, Ichimura, Newey and Powell (1991), Fan (1991), Wansbeek and Meijer (2000), Newey (2001), Taupin (2001), Li (2002), Schennach (2004a, b), Carroll et al. (2004), to name only a few. However, as reviewed in Bound, Brown, and Mathiowetz (2001), validation studies in economic survey data sets indicate that the errors in self-reported variables, such as earnings, are typically correlated with the true values, and hence, nonclassical. In fact, in many survey situations a rational agent has incentive to purposely report wrong values conditioning on his/her truth. Ample empirical evidences of nonclassical measurement errors have drawn growing attentions from theoretical research on econometric models with nonclassical measurement errors. In the meanwhile, given that linear models with measurement errors have been studies thoroughly, recent research activities have been focusing on nonlinear (and/or nonparametric) Errors-In-Variables (EIV) models. However, the identification and estimation of general nonlinear (and/or nonparametric) models with nonclassical errors are notoriously difficult. In this paper, we provide one solution to the nonlinear (and nonparametric) EIV problem by combining a primary sample and an auxiliary sample, where each sample only contains one measurement of the error-ridden variable, and the measurement errors in both samples may be nonclassical. Our identification strategy does not require the existence of instrumental variables for the nonlinear model of interest, nor does it require an auxiliary sample containing the true values nor independence between the two samples.

It is well known that, without additional information or restrictions, a general nonlinear model can not be identified in the presence of measurement errors. When point identification is not feasible under weak assumptions, some research activities have focused on partial identification and bound analyses. See e.g., Chesher (1991), Horowitz and Manski (1995), Manski and Tamer (2003), Molinari (2004) and others.

One approach to regain identification and consistent estimation of nonlinear EIV models with classical errors is to impose parametric restrictions on error distributions; see e.g., Fan (1991), Buzas and Stefanski (1996), Taupin (2001), Hong and Tamer (2003) and others. However, it might be difficult to impose a correct parametric specification on error distributions for a nonlinear EIV model with nonclassical errors.

Another popular approach to identification and estimation of EIV models is to assume the existence of Instrumental Variables (IVs). See Mahajan (2006), Lewbel (2006) and Hu (2006) for using IV approach to obtain identification and consistent estimation of nonlinear models with misclassification errors in discrete explanatory variables. There exist a large amount of important work on using IV approach to solve linear, nonlinear and/or nonparametric EIV models with classical errors in continuous explanatory variables. See, e.g., Amemiya (1985), Amemiya and Fuller (1988), Carroll and Stefanski (1990), Hausman, Ichimura, Newey, and Powell (1991), Hausman, Newey, and Powell (1995), Wang and Hsiao (1995), Li and Vuong (1998), Newey (2001), Li (2002), Schennach (2004a, b), and Carroll, et al. (2004), to name only a few. Most recently, Hu and Schennach (2006) establish identification and estimation of nonlinear EIV models with nonclassical errors in continuous explanatory variables using IVs, where the IVs are excluded from the nonlinear model of interest and are independent of the measurement errors. Although the IV approach is powerful, but it might be difficult to find a valid IV for a general nonlinear EIV model with nonclassical errors in applications.

The alternative popular approach to identify nonlinear EIV models with nonclassical errors is to combine two samples. See Carroll, Ruppert and Stefanski (1995) and Ridder and Moffitt (2006) for detailed survey about this approach. The advantage of this approach is that the primary sample could contain arbitrary measurement errors and no requirement of existence of IVs. However, the earlier works using this approach typically assume the existence of a true validation sample (i.e., an i.i.d. sample from the same population as the primary sample and contains an accurate measurement of the true values). See e.g., Bound, Brown, Duncan, and Rodgers (1989), Hausman, Ichimura, Newey, and Powell (1991), Carroll and Wand (1991), and Lee and Sepanski (1995), to name only a few. Recent works using the two-sample approach have relaxed the true validation sample requirement. For example, Hu and Ridder (2006) show that the marginal distribution of the latent true values from an independent auxiliary sample is enough to identify nonlinear EIV models with a classical error. Chen, Hong, and Tamer (2005), Chen, Hong and Tarozzi (2005), and Ichimura and Martinez-Sanchis (2006) identify and estimate nonlinear EIV models with nonclassical errors using an auxiliary sample, which could be obtained as a stratified sample of the primary sample. Their approach does not require the auxiliary sample to be a true validation sample nor have the same marginal distributions as those of the primary sample. Nevertheless, they still require that the auxiliary sample contains an accurate measurement of the true values; such a sample might be difficult to find in some applications.

In this paper, we provide nonparametric identification of a nonlinear EIV model with measurement errors in covariates by combining a primary sample and an auxiliary sample, where each sample contains only one measurement of the error-ridden explanatory variable,

and the errors in both samples may be nonclassical. Our approach differs from the IV approach in that we do not require an IV excluded from the model of interest and all the variables in our samples may be included in the model. Our approach is closer to the existing two-sample approach since we also require an auxiliary sample and allow for nonclassical measurement errors in both samples. However, our identification strategy differs from the existing two-sample approach because neither of our samples contains an accurate measurement of the true values.

We assume that the primary sample consists of some dependent variables, some error-free covariates and an error-ridden covariate, where the measurement error has unknown distribution and is allowed to be arbitrarily correlated with the latent true values. The auxiliary sample consists of some error-free covariates and another measurement of the mismeasured covariate. Even under the assumption that the measurement error in the primary sample is independent of other variables conditional on the latent true values, it is clear that a general nonlinear EIV model is not identified using the primary sample only, let alone using the auxiliary sample only. The identification is made possible by combining the two samples. We assume there are contrasting subsamples in the primary and the auxiliary samples. These subsamples may be geographic areas, different age groups, or in general subpopulations with different observed demographic characteristics. We use the difference of the marginal distributions of the latent true values in the contrasting subsamples of both the primary sample and the auxiliary sample to show the error distribution is identified. To be specific, assuming that the distributions of the common error-free covariates conditional on the latent true values are the same in the two samples, we may identify the relationship between the measurement error distribution in the auxiliary sample and the ratio of the marginal distribution of latent true values in the subsamples. In fact, the ratio of the marginal distributions plays a role of an eigenvalue of an observed linear operator, while the measurement error distribution in the auxiliary sample is the corresponding eigenfunction. Therefore, the measurement error distribution may be identified through a diagonal decomposition of an observed linear operator under a normalization condition that the measurement error distribution in the auxiliary sample has zero mode (or zero median or mean). The nonlinear model of interest, defined here as the joint distribution of the dependent variables, all the error-free covariates and the latent true covariate in the primary sample, may then be nonparametrically identified. In this paper, we first illustrate our identification strategy using a nonlinear EIV model with nonclassical errors in discrete covariates of two samples. We then focus on nonparametric identification of a general latent nonlinear model with arbitrary measurement errors in continuous covariates.

Our identification result allows for fully nonparametric EIV models and allows for cor-

related two samples. But, in most empirical applications, the latent models of interest are parametric nonlinear models and the two samples are regarded as independent. Within this framework, we propose a sieve quasi maximum likelihood estimation (MLE) for the latent nonlinear model of interest using two samples with nonclassical measurement errors. Under possible misspecification of the latent parametric model, we establish root-n consistency and asymptotic normality of the sieve quasi MLE of the finite dimensional parameter of interest, as well as its semiparametric efficiency under correct specification. However, different economic models typically imply different parametrically specified structural econometric models, and parametric nonlinear models could be all misspecified. We then provide a sieve likelihood ratio model selection test to compare two possibly misspecified parametric nonlinear EIV models using two independent samples with arbitrary errors. These results are extensions of those in White (1982) and Vuong (1989) to possibly misspecified latent parametric nonlinear structural models, and are also applicable to other possibly misspecified semiparametric models involving unobserved heterogeneity and/or nonparametric endogeneity. For example, one could apply these results to derive valid inference without imposing the correct specification of the parametric structural model in the famous mixture model of Heckman and Singer (1984).

Finally, we present a small Monte Carlo simulation and an empirical illustration. We first use simulated data to estimate a probit model with different nonclassical measurement errors. The Monte Carlo simulations show that the new two-sample sieve MLE performs well with the simulated data. Second, we apply our new estimator to a probit model to estimate the effect of earnings on the voting behavior. It is well known that self-reported earnings contains nonclassical errors. The primary sample is from the Current Population Survey (CPS) in November 2004 and the auxiliary sample is from Survey of Income and Program Participation (SIPP). We use different marital status and gender as contrasting subsamples to identify the error distributions in the auxiliary sample. This empirical illustration shows that our new estimator performs well with real data.

The rest of the paper is organized as follows. Section 2 establishes the nonparametric identification of a general nonlinear EIV model with (possibly) nonclassical errors using two samples. Section 3 presents the two-sample sieve quasi MLE and the sieve likelihood ratio model selection test under possibly misspecified parametric latent models. Section 4 applies the two-sample sieve MLE to a latent probit model with simulated data and real data. Section 5 briefly concludes, and the Appendix contains the proofs of the large sample properties of the sieve quasi MLEs.

2 Nonparametric Identification

2.1 The dichotomous case: an illustration

We first illustrate our identification strategy in the special case where the key variables in the model are 0-1 dichotomous. Suppose that we are interested in the effect of the true college education level X^* on the labor supply Y with the marital status W^u and the gender W^v as covariates. This effect would be identified if we could identify the joint density $f_{X^*,W^u,W^v,Y}$. In this example, we assume X^*, W^u, W^v are all 0-1 dichotomous. The true education level X^* is unobserved and subject to measurement errors, (W^u, W^v) are accurately measured and observed in both the primary sample and the auxiliary sample, and Y is only observed in the primary sample. The primary sample is a random sample from (X, W^u, W^v, Y), where X is a mismeasured X^*. In the auxiliary sample, we observe (X_a, W_a^u, W_a^v), where the observed X_a is a proxy of a latent education level X_a^*, W_a^u is the marital status, and W_a^v is the gender. In this illustration subsection, we use italic letters to highlight all the assumptions imposed for the nonparametric identification of $f_{X^*,W^u,W^v,Y}$, while detailed discussions of the assumptions are postponed to subsection 2.2.

We assume that *the measurement error in X is independent of all other variables in the model conditional on the true value X^*, i.e.,* $f_{X|X^*,W^u,W^v,Y} = f_{X|X^*}$. In this simple example, this assumption implies that all the people with the same education level have the same pattern of misreporting the latent true education level, which can be relaxed if there are more common covariates in the two samples. Under this assumption, the probability distribution of the observables equals

$$f_{X,W^u,W^v,Y}(x,u,v,y) = \sum_{x^*=0,1} f_{X|X^*}(x|x^*) f_{X^*,W^u,W^v,Y}(x^*,u,v,y) \quad \text{for all } x,u,v,y. \tag{2.1}$$

We define the matrix representations of $f_{X|X^*}$ as follows:

$$L_{X|X^*} = \begin{pmatrix} f_{X|X^*}(0|0) & f_{X|X^*}(0|1) \\ f_{X|X^*}(1|0) & f_{X|X^*}(1|1) \end{pmatrix}.$$

Notice that the matrix $L_{X|X^*}$ contains the same information as the conditional density $f_{X|X^*}$.

Equation (2.1) then implies for all u, v, y

$$\begin{pmatrix} f_{X,W^u,W^v,Y}(0,u,v,y) \\ f_{X,W^u,W^v,Y}(1,u,v,y) \end{pmatrix} = L_{X|X^*} \times \begin{pmatrix} f_{X^*,W^u,W^v,Y}(0,u,v,y) \\ f_{X^*,W^u,W^v,Y}(1,u,v,y) \end{pmatrix}. \tag{2.2}$$

Equation (2.2) implies that the density $f_{X^*,W^u,W^v,Y}$ would be identified provided that $L_{X|X^*}$ would be identifiable and invertible. Moreover, equation (2.1) implies for the subsamples of males ($W^v = 1$) and of females ($W^v = 0$)

$$\begin{aligned} f_{X,W^u|W^v=j}(x,u) &= \sum_{x^*=0,1} f_{X|X^*,W^u,W^v=j}(x|x^*,u)\, f_{W^u|X^*,W^v=j}(u|x^*) f_{X^*|W^v=j}(x^*). \\ &= \sum_{x^*=0,1} f_{X|X^*}(x|x^*)\, f_{W^u|X^*,W^v=j}(u|x^*) f_{X^*|W^v=j}(x^*), \end{aligned} \tag{2.3}$$

where $f_{X,W^u|W^v=j}(x,u) \equiv f_{X,W^u|W^v}(x,u|j)$ and $j = 0, 1$. By counting the numbers of knows and unknowns in equation (2.3), one can see that the unknown density $f_{X|X^*}$ together with other unknowns can not be identified using the primary sample alone.

In the auxiliary sample, we assume that *the measurement error in X_a satisfies the same conditional independence assumption as that in X, i.e., $f_{X_a|X_a^*,W_a^u,W_a^v} = f_{X_a|X_a^*}$*. Furthermore, we link the two samples by a stable assumption that *the distribution of the marital status conditional on the true education level and gender is the same in the two samples, i.e., $f_{W_a^u|X_a^*,W_a^v=j}(u|x^*) = f_{W^u|X^*,W^v=j}(u|x^*)$ for all u, j, x^**. Therefore, we have for the subsamples of males ($W_a^v = 1$) and of females ($W_a^v = 0$)

$$\begin{aligned} f_{X_a,W_a^u|W_a^v=j}(x,u) &= \sum_{x^*=0,1} f_{X_a|X_a^*,W_a^u,W_a^v=j}(x|x^*,u)\, f_{W_a^u|X_a^*,W_a^v=j}(u|x^*) f_{X_a^*|W_a^v=j}(x^*) \\ &= \sum_{x^*=0,1} f_{X_a|X_a^*}(x|x^*)\, f_{W^u|X^*,W^v=j}(u|x^*) f_{X_a^*|W_a^v=j}(x^*). \end{aligned} \tag{2.4}$$

We define the matrix representations of relevant densities for the subsamples of males

$(W^v = 1)$ and of females $(W^v = 0)$ in the primary sample as follows: for $j = 0, 1$,

$$
\begin{aligned}
L_{X,W^u|W^v=j} &= \begin{pmatrix} f_{X,W^u|W^v=j}(0,0) & f_{X,W^u|W^v=j}(0,1) \\ f_{X,W^u|W^v=j}(1,0) & f_{X,W^u|W^v=j}(1,1) \end{pmatrix} \\
L_{W^u|X^*,W^v=j} &= \begin{pmatrix} f_{W^u|X^*,W^v=j}(0|0) & f_{W^u|X^*,W^v=j}(0|1) \\ f_{W^u|X^*,W^v=j}(1|0) & f_{W^u|X^*,W^v=j}(1|1) \end{pmatrix}^T \\
L_{X^*|W^v=j} &= \begin{pmatrix} f_{X^*|W^v=j}(0) & 0 \\ 0 & f_{X^*|W^v=j}(1) \end{pmatrix},
\end{aligned}
$$

where the superscript T stands for the transpose of a matrix. We similarly define the matrix representations $L_{X_a,W_a^u|W_a^v=j}$, $L_{X_a|X_a^*}$, $L_{W_a^u|X_a^*,W_a^v=j}$ and $L_{X_a^*|W_a^v=j}$ of the corresponding densities $f_{X_a,W_a^u|W_a^v=j}$, $f_{X_a|X_a^*}$, $f_{W_a^u|X_a^*,W_a^v=j}$ and $f_{X_a^*|W_a^v=j}$ in the auxiliary sample. We note that equation (2.3) implies for $j = 0, 1$,

$$
\begin{aligned}
& L_{X|X^*} L_{X^*|W^v=j} L_{W^u|X^*,W^v=j} \\
= \; & L_{X|X^*} \begin{pmatrix} f_{X^*|W^v=j}(0) & 0 \\ 0 & f_{X^*|W^v=j}(1) \end{pmatrix} \begin{pmatrix} f_{W^u|X^*,W^v=j}(0|0) & f_{W^u|X^*,W^v=j}(0|1) \\ f_{W^u|X^*,W^v=j}(1|0) & f_{W^u|X^*,W^v=j}(1|1) \end{pmatrix}^T \\
= \; & L_{X|X^*} \begin{pmatrix} f_{W^u,X^*|W^v=j}(0,0) & f_{W^u,X^*|W^v=j}(1,0) \\ f_{W^u,X^*|W^v=j}(0,1) & f_{W^u,X^*|W^v=j}(1,1) \end{pmatrix} \\
= \; & \begin{pmatrix} f_{X|X^*}(0|0) & f_{X|X^*}(0|1) \\ f_{X|X^*}(1|0) & f_{X|X^*}(1|1) \end{pmatrix} \begin{pmatrix} f_{W^u,X^*|W^v=j}(0,0) & f_{W^u,X^*|W^v=j}(1,0) \\ f_{W^u,X^*|W^v=j}(0,1) & f_{W^u,X^*|W^v=j}(1,1) \end{pmatrix} \\
= \; & \begin{pmatrix} f_{X,W^u|W^v=j}(0,0) & f_{X,W^u|W^v=j}(0,1) \\ f_{X,W^u|W^v=j}(1,0) & f_{X,W^u|W^v=j}(1,1) \end{pmatrix} \\
= \; & L_{X,W^u|W^v=j} \; ,
\end{aligned}
$$

that is

$$L_{X,W^u|W^v=j} = L_{X|X^*} L_{X^*|W^v=j} L_{W^u|X^*,W^v=j}. \tag{2.5}$$

Similarly, equation (2.4) implies

$$L_{X_a,W_a^u|W_a^v=j} = L_{X_a|X_a^*} L_{X_a^*|W_a^v=j} L_{W^u|X^*,W^v=j}. \tag{2.6}$$

Assuming that *the observable matrices* $L_{X_a,W_a^u|W_a^v=j}$ *and* $L_{X,W^u|W^v=j}$ *are invertible, that the diagonal matrices* $L_{X^*|W^v=j}$ *and* $L_{X_a^*|W_a^v=j}$ *are invertible, and that* $L_{X_a|X_a^*}$ *is invertible.* Then equations (2.5) and (2.6) imply that $L_{X|X^*}$ and $L_{W^u|X^*,W^v=j}$ are invertible, and we can now eliminate $L_{W^u|X^*,W^v=j}$ to have for $j = 0, 1$

$$L_{X_a,W_a^u|W_a^v=j} L^{-1}_{X,W^u|W^v=j} = L_{X_a|X_a^*} L_{X_a^*|W_a^v=j} L^{-1}_{X^*|W^v=j} L^{-1}_{X|X^*}.$$

Since this equation hold for $j = 0, 1$, we may then eliminate $L_{X|X^*}$ to have

$$\begin{aligned} L_{X_a,X_a} &\equiv \left(L_{X_a,W_a^u|W_a^v=1} L^{-1}_{X,W^u|W^v=1}\right)\left(L_{X_a,W_a^u|W_a^v=0} L^{-1}_{X,W^u|W^v=0}\right)^{-1} \\ &= L_{X_a|X_a^*}\left(L_{X_a^*|W_a^v=1} L^{-1}_{X^*|W^v=1} L_{X^*|W^v=0} L^{-1}_{X_a^*|W_a^v=0}\right) L^{-1}_{X_a|X_a^*} \\ &\equiv \begin{pmatrix} f_{X_a|X_a^*}(0|0) & f_{X_a|X_a^*}(0|1) \\ f_{X_a|X_a^*}(1|0) & f_{X_a|X_a^*}(1|1) \end{pmatrix} \begin{pmatrix} k_{X_a^*}(0) & 0 \\ 0 & k_{X_a^*}(1) \end{pmatrix} \times \\ &\quad \times \begin{pmatrix} f_{X_a|X_a^*}(0|0) & f_{X_a|X_a^*}(0|1) \\ f_{X_a|X_a^*}(1|0) & f_{X_a|X_a^*}(1|1) \end{pmatrix}^{-1}. \end{aligned} \tag{2.7}$$

with

$$k_{X_a^*}(x^*) = \frac{f_{X_a^*|W_a^v=1}(x^*) f_{X^*|W^v=0}(x^*)}{f_{X^*|W^v=1}(x^*) f_{X_a^*|W_a^v=0}(x^*)}.$$

Notice that the matrix $\left(L_{X_a^*|W_a^v=1} L^{-1}_{X^*|W^v=1} L_{X^*|W^v=0} L^{-1}_{X_a^*|W_a^v=0}\right)$ is diagonal because $L_{X^*|W^v=j}$ and $L_{X_a^*|W_a^v=j}$ are diagonal matrices. The equation (2.7) provides an eigenvalue-eigenvector decomposition of an observed matrix L_{X_a,X_a} on the left-hand side. If such a decomposition is unique, then we may identify $L_{X_a|X_a^*}$, i.e., $f_{X_a|X_a^*}$, from the observed matrix L_{X_a,X_a}.

We assume that $k_{X_a^*}(0) \neq k_{X_a^*}(1)$, *i.e., the eigenvalues are distinctive.* This assumption

requires that the distributions of the latent education level of males or females in the primary sample are different from those in the auxiliary sample, and that the distribution of the latent education level of males is different from that of females in one of the two samples. Notice that each eigenvector is a column in $L_{X_a|X_a^*}$, which is a conditional density. That means each eigenvector is automatically normalized. Therefore, for an observed L_{X_a,X_a}, we may have an eigenvalue-eigenvector decomposition as follows:

$$L_{X_a,X_a} = \begin{pmatrix} f_{X_a|X_a^*}(0|x_1^*) & f_{X_a|X_a^*}(0|x_2^*) \\ f_{X_a|X_a^*}(1|x_1^*) & f_{X_a|X_a^*}(1|x_2^*) \end{pmatrix} \begin{pmatrix} k_{X_a^*}(x_1^*) & 0 \\ 0 & k_{X_a^*}(x_2^*) \end{pmatrix} \times \tag{2.8}$$
$$\times \begin{pmatrix} f_{X_a|X_a^*}(0|x_1^*) & f_{X_a|X_a^*}(0|x_2^*) \\ f_{X_a|X_a^*}(1|x_1^*) & f_{X_a|X_a^*}(1|x_2^*) \end{pmatrix}^{-1}.$$

The value in each entry on the right-hand side of equation (2.8) can be directly computed from the observed matrix L_{X_a,X_a}. The only ambiguity left in equation (2.8) is the value of the indices x_1^* and x_2^*, or the indexing of the eigenvalues and eigenvectors. In other words, the identification of $f_{X_a|X_a^*}$ boils down to finding a 1-to-1 mapping between the following two sets of indices of the eigenvalues and eigenvectors:

$$\{x_1^*, x_2^*\} \Longleftrightarrow \{0, 1\}.$$

Next, we make a normalization assumption that *people with (or without) college education in the auxiliary sample are more likely to report that they have (or do not have) college education, i.e.,* $f_{X_a|X_a^*}(x^*|x^*) > 0.5$ *for* $x^* = 0, 1$. (This assumption also implies the invertibility of $L_{X_a|X_a^*}$.) Since the values of $f_{X_a|X_a^*}(0|x_1^*)$ and $f_{X_a|X_a^*}(1|x_1^*)$ are known in equation (2.8), this assumption pins down the index x_1^* as follows:

$$x_1^* = \begin{cases} 0 & \text{if } f_{X_a|X_a^*}(0|x_1^*) > 0.5 \\ 1 & \text{if } f_{X_a|X_a^*}(1|x_1^*) > 0.5 \end{cases}.$$

The value of x_2^* may be found in the same way. In summary, we have identified $L_{X_a|X_a^*}$, i.e.,

$f_{X_a|X_a^*}$, from the decomposition of the observed matrix L_{X_a,X_a}.

After identifying $L_{X_a|X_a^*}$, we can identify $L_{W^u|X^*,W^v=j}$ or $f_{W^u|X^*,W^v=j}$ from equation (2.6) as follows:

$$L_{X_a^*|W_a^v=j} L_{W^u|X^*,W^v=j} = L^{-1}_{X_a|X_a^*} L_{X_a,W_a^u|W_a^v=j},$$

where two matrices $L_{X_a^*|W_a^v=j}$ and $L_{W^u|X^*,W^v=j}$ can be identified through their product on the left-hand side. Moreover, the density $f_{X|X^*}$ or the matrix $L_{X|X^*}$ is identified from equation (2.5) as follows:

$$L_{X|X^*} L_{X^*|W^v=j} = L_{X,W^u|W^v=j} L^{-1}_{W^u|X^*,W^v=j},$$

where we may identify two matrices $L_{X|X^*}$ and $L_{X^*|W^v=j}$ from their product on the left-hand side. Finally, the density of interest $f_{X^*,W^u,W^v,Y}$ is identified from equation (2.2).

This simple example with dichotomous variables demonstrates that we can nonparametrically identify the model of interest using the similarity of the error structures and the difference in the latent distributions between the two samples. We next show that such a nonparametric identification strategy is in fact generally applicable.

2.2 The general case

We are interested in a model containing variables X^*, W, and Y. We say the model is identified if we can identify the joint probability density of X^*, W, Y:

$$f_{X^*,W,Y}(x^*, w, y), \tag{2.9}$$

where X^* is an unobserved scalar covariate subject to measurement errors, W is a vector of accurately measured covariates that are observed in both the primary sample and the auxiliary sample, and Y is a vector of other variables, including dependent variables and other covariates, which are observed in the primary sample only. The primary sample is a random sample from (X, W^T, Y^T), where X is a mismeasured X^*. Suppose the supports of X, W, Y and X^* are $\mathcal{X} \subseteq \mathbb{R}$, $\mathcal{W} \subseteq \mathbb{R}^{d_w}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, and $\mathcal{X}^* \subseteq \mathbb{R}$, respectively. Let $f_{X|X^*}$ and $f_{X^*|X}$ denote the conditional densities of X given X^* and of X^* given X respectively. Let f_X and f_{X^*} denote the marginal densities of X and X^* respectively. We assume that the measurement error in X satisfies

Assumption 2.1 *$f_{X|X^*,W,Y}(x|x^*, w, y) = f_{X|X^*}(x|x^*)$ for all $x \in \mathcal{X}$, $x^* \in \mathcal{X}^*$, $w \in \mathcal{W}$, and $y \in \mathcal{Y}$.*

Assumption 2.1 implies that the measurement error in X is independent of all other variables in the model conditional on the true value X^*. The measurement error in X may still be correlated with the true value X^* in an arbitrary way, and hence, is nonclassical. We realize that assumption 2.1 might be restrictive in some applications. But it is reasonable to believe that the latent true value X^* is a more important factor in the reported value X than any other variables W and Y.

Assumption 2.1 allows for a very general nonclassical error and captures the major concern on misreporting errors. Under assumption 2.1, the probability density of the observed vectors equals

$$f_{X,W,Y}(x,w,y) = \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*) f_{X^*,W,Y}(x^*,w,y)dx^* \quad \text{for all } x,w,y. \tag{2.10}$$

Let $\mathcal{L}^p(\mathcal{X})$, $1 \leq p < \infty$ denote the space of functions with $\int_{\mathcal{X}} |h(x)|^p dx < \infty$, and $\mathcal{L}^\infty(\mathcal{X})$ be the space of functions with $\sup_{x\in\mathcal{X}} |h(x)| < \infty$. Then it is clear that for any fixed $w \in \mathcal{W}$, $y \in \mathcal{Y}$, $f_{X,W,Y}(\cdot,w,y) \in \mathcal{L}^p(\mathcal{X})$ and $f_{X^*,W,Y}(\cdot,w,y) \in \mathcal{L}^p(\mathcal{X}^*)$ for all $1 \leq p \leq \infty$. Let $\mathcal{H}_X \subseteq \mathcal{L}^2(\mathcal{X})$ and $\mathcal{H}_{X^*} \subseteq \mathcal{L}^2(\mathcal{X}^*)$. We define the integral operator $L_{X|X^*} : \mathcal{H}_{X^*} \to \mathcal{H}_X$ as follows:

$$\{L_{X|X^*}h\}(x) = \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*)\, h(x^*)\, dx^* \quad \text{for any } h \in \mathcal{H}_{X^*},\ x \in \mathcal{X}.$$

Therefore, equation (2.10) becomes: $f_{X,W,Y}(x,w,y) = \{L_{X|X^*} f_{X^*,W,Y}(\cdot,w,y)\}(x)$. Then the latent density $f_{X^*,W,Y}$ would be identified from the observed density $f_{X,W,Y}$ provided that the operator $L_{X|X^*}$ would be identifiable and invertible. We will show that $f_{X|X^*}$ can be identified by combining the information of the primary sample with an auxiliary sample.

Suppose that we observe an auxiliary sample, which is a random sample from (X_a, W_a^T), where X_a is a mismeasured X_a^*. In order to combine the two samples they should have something in common. We consider a series of mutually exclusive subsets $V_1, V_2, ..., V_J \subset \mathcal{W}$ in the two samples. For example, the two samples may contain subpopulations with different demographic characteristics, such as, race, gender, profession, and geographic locations. Suppose $W = (W^u, W^v)^T$ and $W_a = (W_a^u, W_a^v)^T$, where W^u and W_a^u are scalar covariate with support $\mathcal{W}^u \subseteq \mathbb{R}$, and W^v and W_a^v are discrete variables with the same support $\mathcal{W}^v = \{v_1, v_2, ..., v_J\}$ indicating the characteristics above. We will discuss the case where there exist extra common covariates, i.e., $(W^u, W^v) \subseteq W$ and $(W_a^u, W_a^v) \subseteq W_a$ later in Remark 2.5. We may let $V_j = \{v_j\}$. Let $\mathcal{X}_a \subseteq \mathbb{R}$ denote the support of X_a. We assume

Assumption 2.2 *(i) X_a^*, W_a^u and W_a^v have the same supports as X^*, W^u and W^v respectively; (ii) $f_{X_a|X_a^*,W_a^u,W_a^v}(x|x^*,u,v) = f_{X_a|X_a^*}(x|x^*)$ for all $x \in \mathcal{X}_a$, $x^* \in \mathcal{X}^*$, $u \in \mathcal{W}^u$ and $v \in \mathcal{W}^v$.*

Assumption 2.2 implies that the distribution of measurement error in X_a is independent of (W_a^u, W_a^v) conditional on the true value X_a^*. This assumption is consistent with assumption 2.1 imposed on the primary sample.

Assumption 2.3 *$f_{W_a^u|X_a^*,V_j}(u|x^*) = f_{W^u|X^*,V_j}(u|x^*)$ for all $u \in \mathcal{W}^u \subseteq \mathbb{R}$ and $x^* \in \mathcal{X}^*$.*

Assumption 2.3 implies the conditional distribution of the scalar covariate W^u given the true value X^* is the same in each subsample corresponding to V_j in the two samples. If the conditional densities describe an unknown economic relationship between the two variables, assumption 2.3 requires that such a relationship is stable across the two samples. If such a common covariate does not exist in either of the two samples, there is basically no common information to link them. A sufficient condition for assumption 2.3 is that $f_{W_a|X_a^*} = f_{W|X^*}$. In the case where V_j corresponds to each possible value of W^v, assumption 2.3 can be written as $f_{W_a^u|X_a^*,W_a^v} = f_{W^u|X^*,W^v}$. We note that under assumption 2.3, the marginal distributions of the true value X^* and the vector of covariates W in the primary sample may still be different from those of X_a^* and W_a in the auxiliary sample.

For each subsample V_j, assumption 2.1 implies that

$$\begin{aligned} f_{X,W^u|V_j}(x,u) &= \int f_{X|X^*,W^u,V_j}(x|x^*,u)\, f_{W^u|X^*,V_j}(u|x^*) f_{X^*|V_j}(x^*)dx^*. \\ &= \int f_{X|X^*}(x|x^*)\, f_{W^u|X^*,V_j}(u|x^*) f_{X^*|V_j}(x^*)dx^*. \end{aligned} \tag{2.11}$$

in the primary sample. Similarly, assumptions 2.2 and 2.3 imply that

$$\begin{aligned} f_{X_a,W_a^u|V_j}(x,u) &= \int f_{X_a|X_a^*,W_a^u,V_j}(x|x^*,u)\, f_{W_a^u|X_a^*,V_j}(u|x^*) f_{X_a^*|V_j}(x^*)dx^* \\ &= \int f_{X_a|X_a^*}(x|x^*)\, f_{W^u|X^*,V_j}(u|x^*) f_{X_a^*|V_j}(x^*)dx^* \end{aligned} \tag{2.12}$$

in the auxiliary sample. We define the following operators for the primary sample

$$\left(L_{X,W^u|V_j}h\right)(x) = \int f_{X,W^u|V_j}(x,u)h(u)\,du, \quad \left(L_{W^u|X,V_j}h\right)(x) = \int f_{W^u|X,V_j}(u|x)h(u)\,du,$$

$$\left(L_{W^u|X^*,V_j}h\right)(x^*) = \int f_{W^u|X^*,V_j}(u|x^*)h(u)\,du, \quad \left(L_{X^*|V_j}h\right)(x^*) = f_{X^*|V_j}(x^*)h(x^*).$$

We also define the operators $L_{X_a|X_a^*}$, $L_{X_a,W_a^u|V_j}$, $L_{W_a^u|X_a,V_j}$ and $L_{X_a^*|V_j}$ for the auxiliary sample in the same way as their counterparts in the primary sample. Notice that operators $L_{X^*|V_j}$ and $L_{X_a^*|V_j}$ are diagonal operators. By equation (2.11) and the definition of the operators, we have for any function h,

$$\begin{aligned}
\left(L_{X,W^u|V_j}h\right)(x) &= \int f_{X,W^u|V_j}(x,u)h(u)\,du \\
&= \int \left(\int f_{X|X^*}(x|x^*)\, f_{W^u|X^*,V_j}(u|x^*) f_{X^*|V_j}(x^*)dx^*\right) h(u)\,du \\
&= \int f_{X|X^*}(x|x^*)\, f_{X^*|V_j}(x^*) \left(\int f_{W^u|X^*,V_j}(u|x^*)h(u)\,du\right) dx^* \\
&= \int f_{X|X^*}(x|x^*)\, f_{X^*|V_j}(x^*) \left(L_{W^u|X^*,V_j}h\right)(x^*)\,dx^* \\
&= \int f_{X|X^*}(x|x^*) \left(L_{X^*|V_j}L_{W^u|X^*,V_j}h\right)(x^*)\,dx^* \\
&= \left(L_{X|X^*}L_{X^*|V_j}L_{W^u|X^*,V_j}h\right)(x).
\end{aligned}$$

This means we have the following operator equivalence

$$L_{X,W^u|V_j} = L_{X|X^*}L_{X^*|V_j}L_{W^u|X^*,V_j} \tag{2.13}$$

in the primary sample. Similarly, equation (2.12) and the definition of the operators imply

$$L_{X_a,W_a^u|V_j} = L_{X_a|X_a^*}L_{X_a^*|V_j}L_{W^u|X^*,V_j} \tag{2.14}$$

in the auxiliary sample. While the left-hand sides of equations (2.13) and (2.14) are observed, the right-hand sides contain unknown operators corresponding to the error distributions ($L_{X|X^*}$ and $L_{X_a|X_a^*}$), the marginal distributions of the latent true values ($L_{X^*|V_j}$ and $L_{X_a^*|V_j}$), and the conditional distribution of the scalar common covariate ($L_{W^u|X^*,V_j}$).

Equations (2.13) and (2.14) imply that one can not apply the identification results in Hu and Schennach (2006, the IV approach) to the primary sample (or the auxiliary sample) to identify the error distribution $f_{X|X^*}$ (or $f_{X_a|X_a^*}$). Although the dependent variable in their paper may not be a variable of interest, the IV approach still requires that, conditional on

the latent true values, the dependent variable is independent of the mismeasured values and the IV, and that the dependent variable has to vary with the latent true values. Intuitively, this requirement in the IV approach implies that the dependent variable still contains information on the latent true variable even conditional on all other observed variables, and that the dependent variable can not be generated from the observed variables by researchers. Therefore, the indicator of subsample V_j in our approach can not play the role of dependent variable in their IV approach because V_j is generated from the observed W (or W_a) hence in general W^u (or W_a^u) are not independent of V_j conditional on X^* (or X_a^*). In other words, the common variable W^u (or W_a^u) can not play the role of their IV because it is generally correlated with the indicator of subsample V_j.

In order to identify the unknown operators in equations (2.13) and (2.14), we assume

Assumption 2.4 *$L_{X_a|X_a^*} : \mathcal{H}_{X_a^*} \to \mathcal{H}_{X_a}$ is injective, i.e., the set $\{h \in \mathcal{H}_{X_a^*} : L_{X_a|X_a^*} h = 0\} = \{0\}$.*

Assumption 2.4 implies that the inverse of the linear operator $L_{X_a|X_a^*}$ exists. Recall that the conditional expectation operator of X_a^* given X_a, $E_{X_a^*|X_a} : \mathcal{L}^2(\mathcal{X}^*) \to \mathcal{L}^2(\mathcal{X}_a)$, is defined as

$$\{E_{X_a^*|X_a} h'\}(x) = \int_{\mathcal{X}^*} f_{X_a^*|X_a}(x^*|x)\, h'(x^*) dx^* = E[h'(X_a^*)\,|X_a = x] \text{ for any } h' \in \mathcal{L}^2(\mathcal{X}^*),\ x \in \mathcal{X}_a.$$

We have $\{L_{X_a|X_a^*} h\}(x) = \int_{\mathcal{X}^*} f_{X_a|X_a^*}(x|x^*)\, h(x^*)\, dx^* = E\left[\frac{f_{X_a}(x) h(X_a^*)}{f_{X_a^*}(X_a^*)}|X_a = x\right]$ for any $h \in \mathcal{H}_{X_a^*}$, $x \in \mathcal{X}_a$. Assumption 2.4 is equivalent to: $E\left[h(X_a^*)\frac{f_{X_a}(X_a)}{f_{X_a^*}(X_a^*)}|X_a\right] = 0$ implies $h = 0$. If $0 < f_{X_a^*}(x^*) < \infty$ over $int(\mathcal{X}^*)$ and $0 < f_{X_a}(x) < \infty$ over $int(\mathcal{X}_a)$ (which are very minor restrictions), then assumption 2.4 is the same as the identification condition imposed in Newey and Powell (2003), Darolles, Florens and Renault (2005), Carrasco, Florens, and Renault (2006) and others. Moreover, as shown in Newey and Powell (2003), this condition is implied by the *completeness* of the conditional density $f_{X_a^*|X_a}$, which is satisfied when $f_{X_a^*|X_a}$ belongs to an exponential family. In fact, if we are willing to assume $\sup_{x^*,w} f_{X_a^*,W_a}(x^*, w) \leq c < \infty$, then a sufficient condition for assumption 2.4 is the *bounded completeness* of the conditional density $f_{X_a^*|X_a}$; see e.g., Blundell, Chen and Kristensen (2004) and Chernozhukov, Imbens and Newey (2006). When X_a and X_a^* are discrete, assumption 2.4 requires that the support of X_a is not smaller than that of X_a^*.

Assumption 2.5 *(i) $f_{X^*|V_j} > 0$ and $f_{X_a^*|V_j} > 0$; (ii) $L_{W^u|X,V_j}$ is injective and $f_{X|V_j} > 0$; (iii) $L_{W_a^u|X_a,V_j}$ is injective and $f_{X_a|V_j} > 0$.*

By the definition of $L_{X,W^u|V_j}$, we have $\left(L_{W^u|X,V_j}h\right)(x) = \frac{1}{f_{X|V_j}(x)}\left(L_{X,W^u|V_j}h\right)(x)$. Assumption 2.5(ii) then implies that $L_{X,W^u|V_j}$ is invertible. Similarly, assumption 2.5(iii) implies that $L_{X_a,W_a^u|V_j}$ is invertible. Therefore, assumptions 2.4 and 2.5 imply that all the operators involved in equations (2.13) and (2.14) are invertible. More precisely, assumptions 2.4, 2.5(i) and 2.5(iii) and equation (2.14) imply that the operator $L_{W^u|X^*,V_j}$ is injective. Next, the injectivity of $L_{W^u|X^*,V_j}$, assumptions 2.5(i) and 2.5(ii) and equation (2.13) imply that $L_{X|X^*}$ is injective.

Remark 2.1 *Under equations (2.13) and (2.14) and assumption 2.5 (i), the invertibility (or injectivity) of any three operators from $L_{X,W^u|V_j}$, $L_{X|X^*}$, $L_{W^u|X^*,V_j}$, $L_{X_a,W_a^u|V_j}$ and $L_{X_a|X_a^*}$ implies the invertibility (or injectivity) of the remaining two operators. Therefore assumptions 2.4 and 2.5 (ii)(iii) could be replaced by alternative conditions that still imply the invertibility of all the five operators. We decide to impose assumptions 2.5 (ii) and (iii) since they are conditions on observables only and can be verified from data. Nevertheless, we could replace assumption 2.4 by the assumption that the operator $L_{X|X^*} : \mathcal{H}_{X^*} \to \mathcal{H}_X$ is injective. In this paper we impose assumption 2.4 and allow the conditional distribution $f_{X|X^*}$ in the primary sample to be very flexible.*

Under assumptions 2.4 and 2.5, for any given V_j we can eliminate $L_{W^u|X^*,V_j}$ in equations (2.13) and (2.14) to obtain

$$L_{X_a,W_a^u|V_j}L_{X,W^u|V_j}^{-1} = L_{X_a|X_a^*}L_{X_a^*|V_j}L_{X^*|V_j}^{-1}L_{X|X^*}^{-1}. \tag{2.15}$$

This equation holds for all V_i and V_j. We may then eliminate $L_{X|X^*}$ to have

$$\begin{aligned} L_{X_a,X_a}^{ij} &\equiv \left(L_{X_a,W_a^u|V_j}L_{X,W^u|V_j}^{-1}\right)\left(L_{X_a,W_a^u|V_i}L_{X,W^u|V_i}^{-1}\right)^{-1} \\ &= L_{X_a|X_a^*}\left(L_{X_a^*|V_j}L_{X^*|V_j}^{-1}L_{X^*|V_i}L_{X_a^*|V_i}^{-1}\right)L_{X_a|X_a^*}^{-1} \\ &\equiv L_{X_a|X_a^*}L_{X_a^*}^{ij}L_{X_a|X_a^*}^{-1} \end{aligned} \tag{2.16}$$

The operator L_{X_a,X_a}^{ij} on the left-hand side is observed for all i and j. An important observation is that the operator $L_{X_a^*}^{ij} = \left(L_{X_a^*|V_j}L_{X^*|V_j}^{-1}L_{X^*|V_i}L_{X_a^*|V_i}^{-1}\right)$ is a diagonal operator defined as

$$\left(L_{X_a^*}^{ij}h\right)(x^*) = k_{X_a^*}^{ij}(x^*)\,h(x^*) \tag{2.17}$$

with

$$k_{X_a^*}^{ij}(x^*) \equiv \frac{f_{X_a^*|V_j}(x^*) f_{X^*|V_i}(x^*)}{f_{X^*|V_j}(x^*) f_{X_a^*|V_i}(x^*)}.$$

Equation (2.16) implies a diagonalization of an observed operator L_{X_a,X_a}^{ij}. An eigenvalue of L_{X_a,X_a}^{ij} equals $k_{X_a^*}^{ij}(x^*)$ for a value of x^*, which corresponds to an eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$.

The structure of equation (2.16) is similar to that of equation 8 in Hu and Schennach (2006, the IV approach) in the sense that both equations provide a diagonal decomposition of observed operators, whose eigenfunctions correspond to measurement error distributions. Therefore, the same technique of operator diagonalization is used for the identification of $f_{X_a|X_a^*}$. On the one hand, these results imply that the technique of operator diagonalization is a very powerful tool in the identification of nonclassical measurement error models. On the other hand, the difference between our equation (2.16) and their equation 8 also shows how the identification strategy in our paper differs from the IV approach in Hu and Schennach (2006). An eigenvalue in their IV approach is a value of the latent density of interest, while an eigenvalue in our paper is a value of the ratio of marginal distributions of the latent true values in different subpopulations. Moreover, the eigenvalues in our paper do not degenerate to those in the IV approach, or vice versa. Therefore, although both papers use the operator decomposition technique, our identification strategy is very different from theirs for the IV approach.

Remark 2.2 *We may also eliminate $L_{X_a|X_a^*}$ in equation (2.15) for different V_i and V_j to obtain*

$$\left(L_{X_a,W_a^u|V_i} L_{X,W^u|V_i}^{-1}\right)^{-1} \left(L_{X_a,W_a^u|V_j} L_{X,W^u|V_j}^{-1}\right) = L_{X|X^*} \left(L_{X^*|V_i} L_{X_a^*|V_i}^{-1} L_{X_a^*|V_j} L_{X^*|V_j}^{-1}\right) L_{X|X^*}^{-1}.$$

This equation also provides a diagonalization of an observed operator on the left-hand side. If we impose the same restriction on $f_{X|X^*}$ as those will be introduced on $f_{X_a|X_a^*}$, the same identification procedure of $f_{X_a|X_a^*}$ also applies to $f_{X|X^*}$. In this paper, we impose the restrictions on the error distribution $f_{X_a|X_a^*}$ in the auxiliary sample so that we may consider more general measurement errors in the primary sample.

We now show the identification of $f_{X_a|X_a^*}$ and $k_{X_a^*}^{ij}(x^*)$. First, we require the operator L_{X_a,X_a}^{ij} to be bounded so that the diagonal decomposition may be unique; see e.g., Dunford and Schwartz (1971). Equation (2.16) implies that the operator L_{X_a,X_a}^{ij} has the same spectrum as the diagonal operator $L_{X_a^*}^{ij}$. Since an operator is bounded by the largest element of its spectrum, it is sufficient to assume

Assumption 2.6 $k^{ij}_{X^*_a}(x^*) < \infty$ *for all* $i, j \in \{1, 2, ..., J\}$ *for all* x^*.

Notice that the subsets $V_1, V_2, ..., V_J \subset \mathcal{W}$ do not need to be collectively exhaustive. We may only consider those subsets in $\mathcal{W}$ in which these assumptions are satisfied.

Second, although it implies a diagonalization of the operator $L^{ij}_{X_a,X_a}$, equation (2.16) does not guarantee distinctive eigenvalues. If there exist duplicate eigenvalues, there exist two linearly independent eigenfunctions corresponding to the same eigenvalue. A linear combination of the two eigenfunctions is also an eigenfunction corresponding to the same eigenvalue. Therefore, the eigenfunctions may not be identified in each decomposition corresponding to a pair of i and j. However, such an ambiguity can be eliminated by an important observation that the observed operators $L^{ij}_{X_a,X_a}$ for all i, j share the same eigenfunctions $f_{X_a|X^*_a}(\cdot|x^*)$. In order to distinguish each linearly independent eigenfunction, we assume

Assumption 2.7 *For any* $x^*_1 \neq x^*_2$, *there exist* $i, j \in \{1, 2, ..., J\}$ *such that* $k^{ij}_{X^*_a}(x^*_1) \neq k^{ij}_{X^*_a}(x^*_2)$.

Assumption 2.7 implies that for any two different eigenfunctions $f_{X_a|X^*_a}(\cdot|x^*_1)$ and $f_{X_a|X^*_a}(\cdot|x^*_2)$, one can always find two subsets V_j and V_i such that the two different eigenfunctions correspond to two different eigenvalues $k^{ij}_{X^*_a}(x^*_1)$ and $k^{ij}_{X^*_a}(x^*_2)$, and therefore, are identified. Although there may exist duplicate eigenvalues in each decomposition corresponding to a pair of i and j, this assumption guarantees that each eigenfunction $f_{X_a|X^*_a}(\cdot|x^*)$ is uniquely determined by combining all the information from a series of decompositions of $L^{ij}_{X_a,X_a}$ for $i, j \in \{1, 2, ..., J\}$.

Remark 2.3 ***(1)*** *Assumption 2.7 does not hold if* $f_{X^*|V_j}(x^*) = f_{X^*_a|V_j}(x^*)$ *for all* V_j *and all* $x^* \in \mathcal{X}^*$. *This assumption requires that the two samples are from different populations; one can not use two subsets of a random sample as the primary sample and the auxiliary sample for identification, which is different from that in Chen, Hong and Tamer (2005). Given assumption 2.3 and the invertibility of the operator* $L_{W^u|X^*,V_j}$, *one could check assumption 2.7 from the observed densities* $f_{W^u|V_j}$ *and* $f_{W^u_a|V_j}$. *In particular, if* $f_{W^u|V_j}(u) = f_{W^u_a|V_j}(u)$ *for all* V_j *and all* $u \in \mathcal{W}^u$, *then assumption 2.7 is not satisfied.* ***(2)*** *Assumption 2.7 does not hold if* $f_{X^*|V_j}(x^*) = f_{X^*|V_i}(x^*)$ *and* $f_{X^*_a|V_j}(x^*) = f_{X^*_a|V_i}(x^*)$ *for all* $V_j \neq V_i$ *and all* $x^* \in \mathcal{X}^*$. *This means that the marginal distribution of* X^* *or* X^*_a *should be different in the subsamples corresponding to different* V_j *in at least one of the two samples. For example, if* X^* *or* X^*_a *are earnings and* V_j *corresponds to gender, then assumption 2.7 requires that the earning distribution of males should be different from that of females in one of the sample (either the primary or the auxiliary). Given the invertibility of the operators* $L_{X|X^*}$ *and* $L_{X_a|X^*_a}$, *one could check assumption 2.7 from the observed densities* $f_{X|V_j}$ *and* $f_{X_a|V_j}$. *In particular,*

if $f_{X|V_j}(x) = f_{X|V_i}(x)$ *for all* $V_j \neq V_i$ *and all* $x \in \mathcal{X}$*, then assumption 2.7 requires the existence of an auxiliary sample such that* $f_{X_a|V_j}(X_a) \neq f_{X_a|V_i}(X_a)$ *with positive probability for some* $V_j \neq V_i$*.*

We now provide an example of the marginal distribution of X^* to illustrate that assumptions 2.6 and 2.7 are easily satisfied. Suppose that the distribution of X^* in the primary sample is the standard normal, i.e., $f_{X^*|V_j}(x^*) = \psi(x^*)$ for $j = 1, 2, 3$, where ψ is the probability density function of the standard normal, and that the distribution of X_a^* in the auxiliary sample is for $0 < \sigma < 1$ and $\mu \neq 0$

$$f_{X_a^*|V_j}(x^*) = \begin{cases} \psi(x^*) & \text{for } j = 1 \\ \sigma^{-1}\psi(\sigma^{-1}x^*) & \text{for } j = 2 \\ \psi(x^* - \mu) & \text{for } j = 3 \end{cases} . \tag{2.18}$$

It is obvious that assumption 2.6 is satisfied with

$$k_{X_a^*}^{ij}(x^*) = \begin{cases} \sigma^{-1}\exp\left(-\frac{1-\sigma^{-2}}{2}(x^*)^2\right) & \text{for } i = 1,\ j = 2 \\ \frac{\psi(x^*-\mu)}{\psi(x^*)} & \text{for } i = 1,\ j = 3 \end{cases} . \tag{2.19}$$

For $i = 1$,$j = 2$, any two eigenvalues $k_{X_a^*}^{12}(x_1^*)$ and $k_{X_a^*}^{12}(x_2^*)$ of L_{X_a,X_a}^{12} may be the same if and only if $x_1^* = -x_2^*$. In other words, we can not distinguish the eigenfunctions $f_{X_a|X_a^*}(\cdot|x_1^*)$ and $f_{X_a|X_a^*}(\cdot|x_2^*)$ in the decomposition of L_{X_a,X_a}^{12} if and only if $x_1^* = -x_2^*$. Since $k_{X_a^*}^{ij}(x^*)$ for $i = 1, j = 3$ is not symmetric around zero, the eigenvalues $k_{X_a^*}^{13}(x_1^*)$ and $k_{X_a^*}^{13}(x_2^*)$ of L_{X_a,X_a}^{13} are different for any $x_1^* = -x_2^*$. Notice that the operators L_{X_a,X_a}^{12} and L_{X_a,X_a}^{13} share the same eigenfunctions $f_{X_a|X_a^*}(\cdot|x_1^*)$ and $f_{X_a|X_a^*}(\cdot|x_2^*)$. Therefore, we may distinguish the eigenfunctions $f_{X_a|X_a^*}(\cdot|x_1^*)$ and $f_{X_a|X_a^*}(\cdot|x_2^*)$ with $x_1^* = -x_2^*$ in the decomposition of L_{X_a,X_a}^{13}. By combining the information obtained from the decompositions of L_{X_a,X_a}^{12} and L_{X_a,X_a}^{13}, we can distinguish the eigenfunctions corresponding to any two different values of x^*.

Third, another ambiguity is that for a given value of x^* an eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$ times a constant is still an eigenfunction corresponding to x^*. To eliminate this ambiguity, we need to normalize each eigenfunction. Notice that $f_{X_a|X_a^*}(\cdot|x^*)$ is a conditional probability density for each x^* hence $\int f_{X_a|X_a^*}(x|x^*)\,dx = 1$ for all x^*. This property of conditional density provides a perfect normalization condition.

Fourth, in order to fully identify each eigenfunction, i.e., $f_{X_a|X_a^*}$, we need to identify the exact value of x^* in each eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$. Notice that the eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$ is identified up to the value of x^*. In other words, we have identified a probability density of X_a conditional on $X_a^* = x^*$ with the value of x^* unknown. An intuitive normalization assumption is that the value of x^* is the mean of this identified probability density, i.e., $x^* = \int x f_{X_a|X_a^*}(x|x^*)\,dx$; this assumption implies that the measurement error in the auxiliary sample has zero mean conditional on the latent true values. An alternative normalization assumption is that the value of x^* is the mode of this identified probability density, i.e., $x^* = \arg\max_x f_{X_a|X_a^*}(x|x^*)$; this assumption implies that the error distribution conditional on the latent true values has zero mode. The intuition of this assumption is that people are more willing to report some values close to the latent true values than those far away from the truth. Another normalization assumption may be that the value of x^* is the median of the identified probability density, i.e., $x^* = \inf\left\{x^* : \int_{-\infty}^{x^*} f_{X_a|X_a^*}(x|x^*)\,dx \geq \frac{1}{2}\right\}$; this assumption implies that the error distribution conditional on the latent true values has zero median, and that people have the same probability of overreporting as that of underreporting. Obviously the zero median condition can be generalized to the assumption that the error distribution conditional on the latent true values has a zero quantile. In summary, we use the following general normalizing condition to identify the exact value of x^* for each eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$.

Assumption 2.8 *There exists a known functional M such that $M\left[f_{X_a|X_a^*}(\cdot|x^*)\right] = x^*$ for all $x^* \in \mathcal{X}^*$.*

Assumption 2.8 requires that the support of X_a can not be smaller than that of X_a^*. Recall that in the dichotomous case, assumption 2.8 with zero median or zero mode also implies the invertibility of $L_{X_a|X_a^*}$ (i.e., assumption 2.4). However, this is no longer true even in the general discrete case. For the general discrete case, a comparable sufficient condition for the invertibility of $L_{X_a|X_a^*}$ is strictly diagonal dominance (i.e., the diagonal entries of $L_{X_a|X_a^*}$ are all larger than 0.5), but, assumption 2.8 with zero mode only requires that the diagonal entries of $L_{X_a|X_a^*}$ are the largest in each row, which can not guarantee the invertibility of $L_{X_a|X_a^*}$ when the support of X_a^* contains more than 2 values.

After fully identifying the density function $f_{X_a|X_a^*}$, we now show that the density of interest $f_{X^*,W,Y}$ and $f_{X|X^*}$ are also identified. By equation (2.12), we have $f_{X_a,W_a^u|V_j} = L_{X_a|X_a^*} f_{W_a^u,X_a^*|V_j}$. By the injectivity of operator $L_{X_a|X_a^*}$, the joint density $f_{W_a^u,X_a^*|V_j}$ may be identified as follows:

$$f_{W_a^u,X_a^*|V_j} = L^{-1}_{X_a|X_a^*} f_{X_a,W_a^u|V_j}.$$

Assumption 2.3 implies that $f_{W_a^u|X_a^*,V_j} = f_{W^u|X^*,V_j}$ so that we may identify $f_{W^u|X^*,V_j}$ through

$$f_{W^u|X^*,V_j}(u|x^*) = \frac{f_{W_a^u,X_a^*|V_j}(u,x^*)}{\int f_{W_a^u,X_a^*|V_j}(u,x^*)du} \quad \text{for all } x^* \in \mathcal{X}^*.$$

By equation (2.13) and the injectivity of the identified operator $L_{W^u|X^*,V_j}$, we have

$$L_{X|X^*}L_{X^*|V_j} = L_{X,W^u|V_j}L^{-1}_{W^u|X^*,V_j}. \tag{2.20}$$

The left-hand side of equation (2.20) equals an operator with the kernel function $f_{X,X^*|V_j} \equiv f_{X|X^*}f_{X^*|V_j}$. Since the right-hand side of equation (2.20) has been identified, the kernel $f_{X,X^*|V_j}$ on the left-hand side is also identified. We may then identify $f_{X|X^*}$ through

$$f_{X|X^*}(x|x^*) = \frac{f_{X,X^*|V_j}(x,x^*)}{\int f_{X,X^*|V_j}(x,x^*)dx} \quad \text{for all } x^* \in \mathcal{X}^*.$$

Finally, assumption 2.1 and the injectivity of $L_{X|X^*}$ imply that the density of interest $f_{X^*,W,Y}$ is identified through

$$f_{X^*,W,Y} = L^{-1}_{X|X^*}f_{X,W,Y}.$$

We summarize the identification result in the following theorem:

Theorem 2.4 *Suppose assumptions 2.1-2.8 hold. Then, the densities $f_{X,W,Y}$ and f_{X_a,W_a} uniquely determine $f_{X^*,W,Y}$, $f_{X|X^*}$, and $f_{X_a|X_a^*}$.*

Remark 2.5 ***(1)*** *When there exists extra common covariates in the two samples, we may consider more generally-defined W^u and W_a^u or relax assumptions on the error distributions in the auxiliary sample. On the one hand, this identification theorem still holds when we replace W^u and W_a^u by a scalar measurable function of W and W_a respectively. For example, let g be a known scalar measurable function. Then the identification theorem is still valid when assumptions 2.2, 2.3 and 2.5(ii-iii) hold with $W^u = g(W)$ and $W_a^u = g(W_a)$. On the other hand, we may relax assumptions 2.1 and 2.2(ii) to allow the error distributions to be conditional on the true values and the extra common covariates;* ***(2)*** *The identification theorem does not require that the two samples are independent of each other.*

3 Sieve Quasi Likelihood Estimation and Inference

Our identification result is very general and does not require the two samples to be independent. However, for many applications it is reasonable to assume that there are two random samples $\left\{X_i, W_i^T, Y_i^T\right\}_{i=1}^n$ and $\left\{X_{aj}, W_{aj}^T\right\}_{j=1}^{n_a}$ that are mutually independent.

As shown in Section 2, all the densities $f_{Y|X^*,W}$, $f_{X|X^*}$, $f_{W^u|X^*,W^v}$, $f_{X^*|W^v}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a^v}$ are nonparametrically identified under assumptions 2.1-2.8. Nevertheless, in empirical studies, we typically have either a semiparametric or a parametric specification of the conditional density $f_{Y|X^*,W}$ as the model of interest. In this section, we treat the other densities $f_{X|X^*}$, $f_{W^u|X^*,W^v}$, $f_{X^*|W^v}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a^v}$ as unknown nuisance functions, but consider a parametrically specified conditional density of Y given (X^*, W^T):

$$\{g(y|x^*, w; \theta) : \theta \in \Theta\}, \quad \Theta \text{ a compact subset of } \mathbb{R}^{d_\theta},\ 1 \leq d_\theta < \infty.$$

Define

$$\theta_0 \equiv \arg\max_{\theta \in \Theta} \int [\log g(y|x^*, w; \theta)] f_{Y|X^*,W}(y|x^*, w) dy.$$

The latent parametric model is *correctly specified* if $g(y|x^*, w; \theta_0) = f_{Y|X^*,W}(y|x^*, w)$ for almost all y, x^*, w^T (and θ_0 is called true parameter value); otherwise it is *misspecified* (and θ_0 is called pseudo-true parameter value); see e.g., White (1982).

In this section we provide a root-n consistent and asymptotically normally distributed sieve MLE of θ_0 regardless if the latent parametric model $g(y|x^*, w; \theta)$ is correctly specified or not. When $g(y|x^*, w; \theta)$ is misspecified, the estimator is better to be called the "sieve quasi MLE" instead of "sieve MLE". (In this paper we have used both terminologies since we allow the latent model $g(y|x^*, w; \theta)$ to either correctly or incorrectly specify the true latent conditional density $f_{Y|X^*,W}$.) Under the correct specification of the latent model, we show that the sieve MLE of θ_0 is automatically semiparametrically efficient, and provide a simple consist estimator of its asymptotic variance. In addition, we provide sieve likelihood ratio model selection test of two non-nested parametric specifications of $f_{Y|X^*,W}$ when both could be misspecified.

To simplify notation but without loss of generality, in this section we assume $W^T = (W^u, W^v)$, $W_a^T = (W_a^u, W_a^v)$ with $W^v, W_a^v \in \{v_1, v_2, ..., v_J\}$. We define V_j as the subset of $\mathcal{W}$ with W^v or W_a^v equal to v_j. Also we assume all the variables Y, W^u, X, W_a^u, X_a are scalars, and each has possibly unbounded support (i.e., each could have the whole real line as its support).

3.1 Sieve likelihood estimation under possible misspecification

Let $\alpha_0 \equiv (\theta_0^T, f_{01}, f_{01a}, f_{02}, f_{02a}, f_{03})^T \equiv (\theta_0^T, f_{X|X^*}, f_{X_a|X_a^*}, f_{X^*|W^v}, f_{X_a^*|W_a^v}, f_{W^u|X^*,W^v})^T$ denote the true parameter value, where θ_0 is really "pseudo-true" when the parametric model $g(y|x^*, w; \theta)$ is incorrectly specified for the unknown true density $f_{Y|X^*,W}$. We introduce a dummy random variable S with $S = 1$ indicating primary sample and $S = 0$ indicating auxiliary sample. Then we have a big combined sample

$$\left\{Z_t^T \equiv \left(S_t X_t, S_t W_t^T, S_t Y_t^T, S_t, (1-S_t)X_t, (1-S_t)W_t^T\right)\right\}_{t=1}^{n+n_a}$$

such that $\{X_t, W_t^T, Y_t^T, S_t = 1\}_{t=1}^{n}$ is the primary sample and $\{X_t, W_t^T, S_t = 0\}_{t=n+1}^{n+n_a}$ is the auxiliary sample. Denote $p \equiv \Pr(S_t = 1) \in (0, 1)$. Then the observed joint likelihood for α_0 is

$$\prod_{t=1}^{n+n_a} \left[p \times f(X_t, W_t, Y_t|S_t = 1; \alpha_0)\right]^{S_t} \left[(1-p) \times f(X_t, W_t|S_t = 0; \alpha_0)\right]^{1-S_t},$$

where

$$\begin{aligned} f(X_t, W_t, Y_t|S_t &= 1; \alpha_0) \equiv f_{X,W,Y}(X_t, W_t, Y_t; \theta_0, f_{01}, f_{02}, f_{03}) \\ &= f_{W^v}(W_t^v) \int f_{01}(X_t|x^*) g(Y_t|x^*, W_t; \theta_0) f_{03}(W_t^u|x^*, W_t^v) f_{02}(x^*|W_t^v) dx^*, \end{aligned}$$

$$\begin{aligned} f(X_t, W_t|S_t &= 0; \alpha_0) \equiv f_{X_a,W_a}(X_t, W_t; f_{01a}, f_{02a}, f_{03}) \\ &= f_{W_a^v}(W_t^v) \int f_{01a}(X_t|x_a^*) f_{03}(W_t^u|x_a^*, W_t^v) f_{02a}(x_a^*|W_t^v) dx_a^*. \end{aligned}$$

Before we present a sieve (quasi-) MLE estimator $\widehat{\alpha}$ for α_0, we need to impose some mild smoothness restrictions on the unknown densities. The sieve method allows for unknown functions belonging to many different function spaces such as Sobolev space, Besov space and others; see e.g., Shen and Wong (1994), Chen and Shen (1998). But, for the sake of concreteness and simplicity, we consider the widely used Hölder space of functions. Let $\xi = (\xi_1, \xi_2)^T \in \mathbb{R}^2$, $a = (a_1, a_2)^T$, and $\nabla^a h(\xi) \equiv \frac{\partial^{a_1+a_2} h(\xi_1,\xi_2)}{\partial \xi_1^{a_1} \partial \xi_2^{a_2}}$ denote the $(a_1 + a_2)$-th derivative. Let $\|\cdot\|_E$ denote the Euclidean norm. Let $\mathcal{V} \subseteq \mathbb{R}^2$ and $\underline{\gamma}$ be the largest integer satisfying $\gamma > \underline{\gamma}$. The Hölder space $\Lambda^\gamma(\mathcal{V})$ of order $\gamma > 0$ is a space of functions $h : \mathcal{V} \mapsto \mathbb{R}$ such that the first $\underline{\gamma}$ derivatives are continuous and bounded, and the $\underline{\gamma}$-th derivative are

Hölder continuous with the exponent $\gamma - \underline{\gamma} \in (0, 1]$. The Hölder space $\Lambda^{\gamma}(\mathcal{V})$ becomes a Banach space under the Hölder norm:

$$\|h\|_{\Lambda^{\gamma}} = \max_{a_1+a_2 \leq \underline{\gamma}} \sup_{\xi} |\nabla^a h(\xi)| + \max_{a_1+a_2=\underline{\gamma}} \sup_{\xi \neq \xi'} \frac{|\nabla^a h(\xi) - \nabla^a h(\xi')|}{\left(\|\xi - \xi'\|_E\right)^{\gamma - \underline{\gamma}}} < \infty.$$

We define a Hölder ball as $\Lambda_c^{\gamma}(\mathcal{V}) \equiv \{h \in \Lambda^{\gamma}(\mathcal{V}) : \|h\|_{\Lambda^{\gamma}} \leq c < \infty\}$. Denote

$$\mathcal{F}_1 = \left\{ f_1(\cdot|\cdot) \in \Lambda_c^{\gamma_1}\left(\mathcal{X} \times \mathcal{X}^*\right) : f_1(\cdot|x^*) > 0, \ \int_{\mathcal{X}} f_1(x|x^*)dx = 1 \textit{ for all } x^* \in \mathcal{X}^* \right\},$$

$$\mathcal{F}_{1a} = \left\{ \begin{array}{c} f_{1a}(\cdot|\cdot) \in \Lambda_c^{\gamma_{1a}}\left(\mathcal{X}_a \times \mathcal{X}^*\right) : \textit{ assumptions 2.4, 2.8 hold,} \\ f_{1a}(\cdot|x^*) > 0, \ \int_{\mathcal{X}_a} f_{1a}(x|x^*)dx = 1 \textit{ for all } x^* \in \mathcal{X}^* \end{array} \right\},$$

$$\mathcal{F}_2 = \left\{ \begin{array}{c} f_2\left(\cdot|w^v\right) \in \Lambda_c^{\gamma_2}\left(\mathcal{X}^*\right) : \textit{ assumptions 2.6, 2.7 hold,} \\ f_2\left(\cdot|w^v\right) > 0, \ \int_{\mathcal{X}^*} f_2\left(x^*|w^v\right) dx^* = 1 \textit{ for all } w^v \in \mathcal{W}^v \end{array} \right\},$$

$$\mathcal{F}_3 = \left\{ \begin{array}{c} f_3\left(\cdot|\cdot, w^v\right) \in \Lambda_c^{\gamma_3}\left(\mathcal{W}^u \times \mathcal{X}^*\right) : f_3\left(\cdot|x^*, w^v\right) > 0, \\ \int_{\mathcal{W}^u} f_3\left(w^u|x^*, w^v\right) dw^u = 1 \textit{ for all } x^* \in \mathcal{X}^*, w^v \in \mathcal{W}^v \end{array} \right\}.$$

We impose the following smoothness restrictions on the densities:

Assumption 3.1 *(i) all the assumptions in theorem 2.4 hold; (ii)* $f_{X|X^*}(\cdot|\cdot) \in \mathcal{F}_1$ *with* $\gamma_1 > 1$*; (iii)* $f_{X_a|X_a^*}(\cdot|\cdot) \in \mathcal{F}_{1a}$ *with* $\gamma_{1a} > 1$*; (iv)* $f_{X^*|W^v}\left(\cdot|w^v\right), f_{X_a^*|W_a^v}\left(\cdot|w^v\right) \in \mathcal{F}_2$ *with* $\gamma_2 > 1/2$ *for all* $w^v \in \mathcal{W}^v$*; (v)* $f_{W^u|X^*,W^v}\left(\cdot|\cdot, w^v\right) \in \mathcal{F}_3$ *with* $\gamma_3 > 1$ *for all* $w^v \in \mathcal{W}^v$*.*

Denote $\mathcal{A} = \Theta \times \mathcal{F}_1 \times \mathcal{F}_{1a} \times \mathcal{F}_2 \times \mathcal{F}_2 \times \mathcal{F}_3$ and $\alpha = \left(\theta^T, f_1, f_{1a}, f_2, f_{2a}, f_3\right)^T$. Then the log-joint likelihood for $\alpha \in \mathcal{A}$ is given by:

$$\begin{aligned} & \sum_{t=1}^{n+n_a} \left\{ S_t \ln\left[p \times f(X_t, W_t, Y_t|S_t = 1; \alpha)\right] + (1 - S_t) \ln\left[(1-p) \times f(X_t, W_t|S_t = 0; \alpha)\right] \right\} \\ = \ & n \ln p + n_a \ln(1-p) + \sum_{t=1}^{n+n_a} \ell(Z_t; \alpha), \end{aligned}$$

where

$$\ell(Z_t;\alpha) \equiv S_t \ell_p(Z_t;\theta,f_1,f_2,f_3) + (1-S_t)\ell_a(Z_t;f_{1a},f_{2a},f_3),$$

$$\begin{aligned}\ln f(X_t,W_t,Y_t|S_t &= 1;\alpha) \equiv \ell_p(Z_t;\theta,f_1,f_2,f_3) \\ &= \ln \int f_1(X_t|x^*)g(Y_t|x^*,W_t;\theta)f_3(W_t^u|x^*,W_t^v)f_2(x^*|W_t^v)dx^* + \ln f_{W^v}(W_t^v)\end{aligned}$$

and

$$\begin{aligned}\ln f(X_t,W_t|S_t &= 0;\alpha) \equiv \ell_a(Z_t;f_{1a},f_{2a},f_3) \\ &= \ln \int f_{1a}(X_t|x_a^*)f_3(W_t^u|x_a^*,W_t^v)f_{2a}(x_a^*|W_t^v)dx_a^* + \ln f_{W_a^v}(W_t^v),\end{aligned}$$

Let $E[\cdot]$ denote the expectation with respect to the underlying true data generating process for Z_t. To stress that our combined data set consisting of two samples, sometimes we let $Z_{pi} = (X_i, W_i^u, W_i^v, Y_i)^T$ denote $i-th$ observation in the primary data set, and $Z_{aj} = (X_{aj}, W_{aj}^u, W_{aj}^v)^T$ denote $j-th$ observation in the auxiliary data set. Then

$$\alpha_0 = \arg\sup_{\alpha\in\mathcal{A}} E\left[\ell(Z_t;\alpha)\right] = \arg\sup_{\alpha\in\mathcal{A}} \left[pE\{\ell_p(Z_{pi};\theta,f_1,f_2,f_3)\} + (1-p)E\{\ell_a(Z_{aj};f_{1a},f_{2a},f_3)\}\right].$$

Let $\mathcal{A}_n = \Theta \times \mathcal{F}_1^n \times \mathcal{F}_{1a}^n \times \mathcal{F}_2^n \times \mathcal{F}_2^n \times \mathcal{F}_3^n$ be a sieve space for $\mathcal{A}$, which is a sequence of approximating spaces that are dense in $\mathcal{A}$ under some pseudo-metric. The two-sample sieve quasi- MLE $\widehat{\alpha}_n = \left(\widehat{\theta}^T, \widehat{f}_1, \widehat{f}_{1a}, \widehat{f}_2, \widehat{f}_{2a}, \widehat{f}_3\right)^T \in \mathcal{A}_n$ for $\alpha_0 \in \mathcal{A}$ is defined as:

$$\widehat{\alpha}_n = \arg\max_{\alpha\in\mathcal{A}_n} \sum_{t=1}^{n+n_a} \ell(Z_t;\alpha) = \arg\max_{\alpha\in\mathcal{A}_n} \left[\sum_{i=1}^{n} \ell_p(Z_{pi};\theta,f_1,f_2,f_3) + \sum_{j=1}^{n_a} \ell_a(Z_{aj};f_{1a},f_{2a},f_3)\right]$$

We could apply infinite-dimensional approximating spaces as sieves $\mathcal{F}_j^n$ for $\mathcal{F}_j, j = 1, 1a, 2, 3$. However, in applications, we shall use finite-dimensional sieve spaces since they are easier to implement. For $j = 1, 1a, 2, 3$, let $p_j^{k_{j,n}}(\cdot)$ be a $k_{j,n} \times 1-$vector of known basis functions, such as power series, splines, Fourier series, etc. Then we denote the sieve space

for $\mathcal{F}_1$, $\mathcal{F}_{1a}$, $\mathcal{F}_2$ and $\mathcal{F}_3$ as follows:

$$\mathcal{F}_1^n = \left\{ f_1(x|x^*) = p_1^{k_{1,n}}(x, x^*)^T \beta_1 \in \mathcal{F}_1 \right\},\ \mathcal{F}_{1a}^n = \left\{ f_{1a}(x_a|x_a^*) = p_{1a}^{k_{1a,n}}(x_a, x_a^*)^T \beta_{1a} \in \mathcal{F}_{1a} \right\},$$

$$\mathcal{F}_2^n = \left\{ f_2\left(x^*|w^v\right) = \sum_{j=1}^{J} I\left(w^v = v_j\right) p_2^{k_{2,n}}(x^*)^T \beta_{2j} \in \mathcal{F}_2 \right\},$$

$$\mathcal{F}_3^n = \left\{ f_3\left(w^u|x^*, w^v\right) = \sum_{j=1}^{J} I\left(w^v = v_j\right) p_3^{k_{3,n}}(w^u, x^*)^T \beta_{3j} \in \mathcal{F}_3 \right\}.$$

We now present two concrete examples of sieve bases; see e.g., Newey (1997), Chen and Shen (1998) and Chen (2006) for additional examples. For simplicity we assume $\mathcal{X}$, $\mathcal{X}_a$, $\mathcal{X}^*$, $\mathcal{W}^u$ is $\mathbb{R}$, then we can let $p_2^{k_{2,n}}(\cdot)$ be a $k_{2,n} \times 1-$vector of either Hermite polynomial bases or wavelet spline bases on $\mathbb{R}$; and for $j = 1, 1a, 3$, $p_j^{k_{j,n}}(\cdot,\cdot)$ can be a $k_{j,n} \times 1-$vector of tensor product of either Hermite polynomial bases or wavelet spline bases on $\mathbb{R}^2$.

Hermite polynomials. Hermite polynomial series $\{H_k : k = 1, 2, ...\}$ is an orthonormal basis of $\mathcal{L}^2(\mathbb{R}, \exp\{-x^2\})$. It can be obtained by applying the Gram-Schmidt procedure to the polynomial series $\{x^{k-1} : k = 1, 2, ...\}$ under the inner product $\langle f, g\rangle_\omega = \int_{\mathcal{R}} f(x)g(x)\exp\{-x^2\}dx$. That is, $H_1(x) = 1/\sqrt{\int_{\mathbb{R}} \exp\{-x^2\}dx} = \pi^{-1/4}$, and for all $k \geq 2$,

$$H_k(x) = \frac{x^{k-1} - \sum_{j=1}^{k-1} \left\langle x^{k-1}, H_j \right\rangle_\omega H_j(x)}{\sqrt{\int_{\mathbb{R}} [x^{k-1} - \sum_{j=1}^{k-1} \langle x^{k-1}, H_j\rangle_\omega H_j(x)]^2 \exp\{-x^2\}dx}}.$$

Let $\text{HPol}(k_n)$ denote the space of Hermite polynomials on $\mathbb{R}$ of degree k_n or less:

$$\text{HPol}(k_n) = \left\{ \sum_{k=1}^{k_n+1} a_k H_k(x) \exp\{-\frac{x^2}{2}\},\ x \in \mathbb{R} :\ a_k \in \mathbb{R} \right\}.$$

See e.g. Gallant and Nychka (1987) and Coppejans and Gallant (2002) for properties and applications of the Hermite polynomial sieve.

Cardinal B-spline wavelets. The cardinal B-spline of order $r \geq 1$ is given by

$$B_r(x) = \frac{1}{(r-1)!} \sum_{j=0}^{r} (-1)^j \binom{r}{j} [\max(0, x-j)]^{r-1},$$

which has support $[0, r]$, is symmetric at $r/2$ and is a piecewise polynomial of highest degree $r-1$. It satisfies $B_r(x) \geq 0$, $\sum_{k=-\infty}^{+\infty} B_r(x-k) = 1$ for all $x \in \mathbb{R}$, which is crucial to preserve the shape of the unknown function to be approximated. Its derivative satisfies $\frac{\partial}{\partial x} B_r(x) = B_{r-1}(x) - B_{r-1}(x-1)$. Let $\text{SplWav}(r-1, 2^{k_n})$ denote the space of spline wavelet bases on $\mathbb{R}$:

$$\text{SplWav}(r-1, 2^{k_n}) = \left\{ \sum_{j=-\infty}^{\infty} \alpha_j 2^{k_n/2} B_r(2^{k_n} x - j),\ x \in \mathbb{R} :\ \alpha_k \in \mathbb{R} \right\}.$$

See Chui (1992) and Chen et al. (1997) for properties and applications of the spline wavelet sieve.

3.1.1 Consistency

The consistency of the two-sample sieve (quasi) MLE $\widehat{\alpha}_n$ can be established by applying either lemma A.1 of Newey and Powell (2003) or theorem 3.1 of Chen (2006). First we define a norm on $\mathcal{A}$ as follows:

$$\|\alpha\|_s = \|\theta\|_E + \|f_1\|_{\infty,\omega_1} + \|f_{1a}\|_{\infty,\omega_{1a}} + \|f_2\|_{\infty,\omega_2} + \|f_{2a}\|_{\infty,\omega_2} + \|f_3\|_{\infty,\omega_3}$$

where $\|h\|_{\infty,\omega_j} \equiv \sup_\xi |h(\xi)\omega_j(\xi)|$ with $\omega_j(\xi) = \left(1 + \|\xi\|_E^2\right)^{-\varsigma_j/2}$, $\varsigma_j > 0$ for $j = 1, 1a, 2, 3$. We assume each of $\mathcal{X}$, $\mathcal{X}_a$, $\mathcal{X}^*$, $\mathcal{W}^u$ is $\mathbb{R}$, and

Assumption 3.2 *(i) the primary sample* $\left\{X_i, W_i^T, Y_i^T\right\}_{i=1}^n$ *and the auxiliary sample* $\left\{X_{aj}, W_{aj}^T\right\}_{j=1}^{n_a}$ *are i.i.d and independent of each other. In addition,* $\lim_{n\to\infty} \frac{n}{n+n_a} = p \in (0,1)$*; (ii)* $g(y|x^*, w; \theta)$ *is continuous in* $\theta \in \Theta$*, and* Θ *is a compact subset of* $\mathbb{R}^{d_\theta}$*; (iii)* $\theta_0 \in \Theta$ *is the unique maximizer of* $\int [\log g(y|x^*, w; \theta)] f_{Y|X^*,W}(y|x^*, w) dy$ *over* $\theta \in \Theta$*.*

Assumption 3.3 *(i)* $-\infty < E[\ell(Z_t;\alpha_0)] < \infty$, $E[\ell(Z_t;\alpha)]$ *is upper semicontinuous on* $\mathcal{A}$ *under the metric* $\|\cdot\|_s$*; (ii) there are a finite* $\kappa > 0$ *and a random variable* $U(Z_t)$ *with* $E\{U(Z_t)\} < \infty$ *such that* $\sup_{\alpha\in\mathcal{A}_n:\|\alpha-\alpha_0\|_s\leq\delta} |\ell(Z_t;\alpha) - \ell(Z_t;\alpha_0)| \leq \delta^\kappa U(Z_t)$.

Assumption 3.4 *(i)* $p_2^{k_{2,n}}(\cdot)$ *is a* $k_{2,n}\times 1-$*vector of spline wavelet basis functions on* $\mathbb{R}$*, and for* $j = 1, 1a, 3$, $p_j^{k_{j,n}}(\cdot,\cdot)$ *is a* $k_{j,n}\times 1-$*vector of tensor product of spline wavelet basis functions on* $\mathbb{R}^2$*; (ii)* $k_n \equiv \max\{k_{1,n}, k_{1a,n}, k_{2,n}, k_{3,n}\} \to \infty$ *and* $k_n/n \to 0$.

Assumption 3.2(i) is a typical condition used in cross-sectional analyses with two samples; see e.g., Ridder and Moffitt (2006). Assumption 3.2(ii-iii) are typical conditions for parametric (quasi-) MLE of θ_0 if X^* could be observed without error. Assumption 3.3(ii) requires the log density is Hölder continuous under the metric $\|\cdot\|_s$ over the sieve space. The following consistency lemma is a direct application of lemma A.1 of Newey and Powell (2003) or theorem 3.1 (or remark 3.1(4), remark 3.3) of Chen (2006), hence we omit its proof.

Lemma 3.1 *Let* $\widehat{\alpha}_n$ *be the two-sample sieve MLE. Under assumptions 3.1-3.4, we have* $\|\widehat{\alpha}_n - \alpha_0\|_s = o_p(1)$.

3.1.2 Convergence rate under weaker metric

Although the population criterion function $E[\ell(Z_t;\alpha)]$ is continuous with respect to the strong norm $\|\cdot\|_s$, but the reverse is not true. It is easy to check that the metric $\|\alpha-\alpha_0\|_s$ is in general not continuous with respect to the population criterion function difference $E[\ell(Z_t;\alpha_0) - \ell(Z_t;\alpha)]$. This is the so-called ill-posed inverse problem, and hence one could not generally obtain a fast convergence rate $o_p\left(n^{-1/4}\right)$ under the strong norm $\|\cdot\|_s$. See e.g., Linton and Whang (2002), Newey and Powell (2003), Darolles, Florens and Renault (2005), Hall and Horowitz (2005), Florens, Johannes and van Bellegem (2005), Carrasco and Florens (2005), Carrasco, Florens and Renault (2006), Chen (2006), Horowitz and Lee (2006), Gagliardini and Scaillet (2006), Bonhomme and Robin (2006), Hoderlein, Klemela and Mammen (2006) for further discussions about the ill-posed inverse problems.

We now follow the approach in Ai and Chen (2003, 2004), and introduce a pseudo metric $\|\cdot\|_2$ that is weaker than $\|\cdot\|_s$ but is continuous with respect to the population criterion function difference $E[\ell(Z_t;\alpha_0) - \ell(Z_t;\alpha)]$, so that the convergence rate of the sieve quasi MLE would be $o_p\left(n^{-1/4}\right)$ under the weaker pseudo metric $\|\cdot\|_2$, which is usually needed to establish the $\sqrt{n}-$asymptotic normality of any semiparametric estimator of θ_0.

Given Lemma 3.1, we can now restrict our attention to a shrinking $||\cdot||_s-$neighborhood around α_0. Let $\mathcal{A}_{0s} \equiv \{\alpha\in\mathcal{A} : ||\alpha-\alpha_0||_s = o(1), ||\alpha||_s \leq c_0 < c\}$ and $\mathcal{A}_{0sn} \equiv \{\alpha\in\mathcal{A}_n :$

$||\alpha - \alpha_0||_s = o(1), ||\alpha||_s \leq c_0 < c\}$. Then, for the purpose of establishing a convergence rate under a pseudo metric that is weaker than $||\cdot||_s$, we can treat $\mathcal{A}_{0s}$ as the new parameter space and $\mathcal{A}_{0sn}$ as its sieve space, and assume that both $\mathcal{A}_{0s}$ and $\mathcal{A}_{0sn}$ are convex parameter spaces. For any α_1, $\alpha_2 \in \mathcal{A}_{0s}$, we consider a continuous path $\{\alpha(\tau) : \tau \in [0,1]\}$ in $\mathcal{A}_{0s}$ such that $\alpha(0) = \alpha_1$ and $\alpha(1) = \alpha_2$. For simplicity we assume that for any α, $\alpha + v \in \mathcal{A}_{0s}$, $\{\alpha + \tau v : \tau \in [0,1]\}$ is a continuous path in $\mathcal{A}_{0s}$, and that $\ell(Z_t; \alpha + \tau v)$ is twice continuously differentiable at $\tau = 0$ for almost all Z_t and any direction $v \in \mathcal{A}_{0s}$. We define the pathwise first derivative as

$$\frac{d\ell(Z_t;\alpha)}{d\alpha}[v] \equiv \frac{d\ell(Z_t;\alpha+\tau v)}{d\tau}|_{\tau=0} \text{ a.s. } Z_t,$$

and the pathwise second derivative as

$$\frac{d^2\ell(Z_t;\alpha)}{d\alpha d\alpha^T}[v,v] \equiv \frac{d^2\ell(Z_t;\alpha+\tau v)}{d\tau^2}|_{\tau=0} \quad \text{a.s. } Z_t.$$

Following Ai and Chen (2004), for any $\alpha_1, \alpha_2 \in \mathcal{A}_{0s}$, we define a pseudo metric $||\cdot||_2$ as follows:

$$\|\alpha_1 - \alpha_2\|_2 \equiv \sqrt{-E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[\alpha_1-\alpha_2, \alpha_1-\alpha_2]\right)}.$$

We show that $\widehat{\alpha}_n$ converges to α_0 at a rate faster than $n^{-1/4}$ under the pseudo metric $\|\cdot\|_2$ with the following assumptions:

Assumption 3.5 *(i)* $\varsigma_j > \gamma_j$ *for* $j = 1, 1a, 2, 3$*; (ii)* $k_n^{-\gamma} = o([n+n_a]^{-1/4})$ *with* $\gamma \equiv \min\{\gamma_1/2, \gamma_{1a}/2, \gamma_2, \gamma_3/2\} > 1/2$.

Assumption 3.6 *(i)* $\mathcal{A}_{0s}$ *is convex at* α_0 *and* $\theta_0 \in int(\Theta)$*; (ii)* $\ell(Z_t;\alpha)$ *is twice continuously pathwise differentiable with respect to* $\alpha \in \mathcal{A}_{0s}$*, and* $\log g(y|x^*, w; \theta)$ *is twice continuously differentiable at* θ_0.

Assumption 3.7 $\sup_{\widetilde{\alpha}\in\mathcal{A}_{0s}} \sup_{\alpha\in\mathcal{A}_{0sn}} \left|\frac{d\ell(Z_t;\widetilde{\alpha})}{d\alpha}\left[\frac{\alpha-\alpha_0}{\|\alpha-\alpha_0\|_s}\right]\right| \leq U(Z_t)$ *for a random variable* $U(Z_t)$ *with* $E\{[U(Z_t)]^2\} < \infty$.

Assumption 3.8 *(i)* $\sup_{v\in\mathcal{A}_{0s}:||v||_s=1} -E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[v,v]\right) \leq C < \infty$*; (ii) uniformly over* $\widetilde{\alpha} \in \mathcal{A}_{0s}$ *and* $\alpha \in \mathcal{A}_{0sn}$*, we have*

$$-E\left(\frac{d^2\ell(Z_t;\widetilde{\alpha})}{d\alpha d\alpha^T}[\alpha-\alpha_0, \alpha-\alpha_0]\right) = \|\alpha-\alpha_0\|_2^2 \times \{1+o(1)\}.$$

Assumption 3.5 guarantees that the sieve approximation error under the strong norm $||\cdot||_s$ goes to zero faster than $[n+n_a]^{-1/4}$. Assumption 3.6 makes sure that the twice pathwise derivatives are well defined with respect to $\alpha \in \mathcal{A}_{0s}$, hence the pseudo metric $\|\alpha-\alpha_0\|_2$ is well defined on $\mathcal{A}_{0s}$. Assumption 3.7 impose an envelope condition. Assumption 3.8(i) implies that $\|\alpha-\alpha_0\|_2 \leq \sqrt{C}\,\|\alpha-\alpha_0\|_s$ for all $\alpha \in \mathcal{A}_{0s}$. Assumption 3.8(ii) implies that there are positive finite constants C_1 and C_2 such that for all $\alpha \in \mathcal{A}_{0sn}$, $C_1\|\alpha-\alpha_0\|_2^2 \leq E[\ell(Z_t;\alpha_0)-\ell(Z_t;\alpha)] \leq C_2\|\alpha-\alpha_0\|_2^2$, that is, $\|\alpha-\alpha_0\|_2^2$ is equivalent to the Kullback-Leibler discrepancy on the local sieve space $\mathcal{A}_{0sn}$. The following convergence rate theorem is a direct application of theorem 3.2 of Chen (2006) to the local parameter space $\mathcal{A}_{0s}$ and the local sieve space $\mathcal{A}_{0sn}$, hence we omit its proof.

Theorem 3.2 *Under assumptions 3.1-3.8, we have*

$$\|\widehat{\alpha}_n-\alpha_0\|_2 = O_P\left(\max\left\{k_n^{-\gamma},\sqrt{\frac{k_n}{n+n_a}}\right\}\right) = O_P\left([n+n_a]^{\frac{-\gamma}{2\gamma+1}}\right) \text{ if } k_n = O\left([n+n_a]^{\frac{1}{2\gamma+1}}\right).$$

3.1.3 Asymptotic normality under possible misspecification

Following the approach in Ai and Chen (2004), we can derive the asymptotic distribution of the sieve quasi MLE $\widehat{\theta}_n$ regardless whether the latent parametric model $g(y|x^*,w;\theta_0)$ is correctly specified or not. First we define an inner product corresponding to the pseudo metric $\|\cdot\|_2$:

$$\langle v_1,v_2\rangle_2 \equiv -E\left\{\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[v_1,v_2]\right\}.$$

Let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A}-\{\alpha_0\}$ under the metric $\|\cdot\|_2$. Then $\left(\overline{\mathbf{V}},\|\cdot\|_2\right)$ is a Hilbert space and we could represent $\overline{\mathbf{V}} = \mathbb{R}^{d_\theta}\times\overline{\mathcal{U}}$ with $\overline{\mathcal{U}} \equiv \overline{\mathcal{F}_1\times\mathcal{F}_{1a}\times\mathcal{F}_2\times\mathcal{F}_2\times\mathcal{F}_3}-\{(f_{01},f_{01a},f_{02},f_{02a},f_{03})\}$. Let $h=(f_1,f_{1a},f_2,f_{2a},f_3)$ denote all the unknown densities. Then the pathwise first derivative can be written as

$$\begin{aligned}\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\alpha-\alpha_0] &= \frac{d\ell(Z_t;\alpha_0)}{d\theta^T}(\theta-\theta_0)+\frac{d\ell(Z;\alpha_0)}{dh}[h-h_0]\\ &= \left(\frac{d\ell(Z_t;\alpha_0)}{d\theta^T}-\frac{d\ell(Z;\alpha_0)}{dh}[\mu]\right)(\theta-\theta_0),\end{aligned}$$

with $h - h_0 \equiv -\mu \times (\theta - \theta_0)$, and where

$$\begin{aligned}\frac{d\ell(Z;\alpha_0)}{dh}[h-h_0] &= \frac{d\ell(Z;\theta_0, h_0(1-\tau)+\tau h)}{d\tau}|_{\tau=0} \\ &= \frac{d\ell(Z_t;\alpha_0)}{df_1}[f_1 - f_{01}] + \frac{d\ell(Z_t;\alpha_0)}{df_{1a}}[f_{1a} - f_{01a}] + \frac{d\ell(Z_t;\alpha_0)}{df_2}[f_2 - f_{02}] \\ &\quad + \frac{d\ell(Z_t;\alpha_0)}{df_{2a}}[f_{2a} - f_{02a}] + \frac{d\ell(Z_t;\alpha_0)}{df_3}[f_3 - f_{03}].\end{aligned}$$

Note that

$$\begin{aligned}& E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[\alpha-\alpha_0, \alpha-\alpha_0]\right) \\ = \;& (\theta-\theta_0)^T E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\theta d\theta^T}\right)(\theta-\theta_0) + 2(\theta-\theta_0)^T E\left(\frac{d^2\ell(Z;\alpha_0)}{d\theta dh^T}[h-h_0]\right) \\ & + E\left(\frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[h-h_0, h-h_0]\right) \\ = \;& (\theta-\theta_0)^T E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\theta d\theta^T} - 2\frac{d^2\ell(Z;\alpha_0)}{d\theta dh^T}[\mu] + \frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu,\mu]\right)(\theta-\theta_0),\end{aligned}$$

with $h - h_0 \equiv -\mu \times (\theta - \theta_0)$, and where

$$\frac{d^2\ell(Z;\alpha_0)}{d\theta dh^T}[h-h_0] = \frac{d(\partial\ell(Z;\theta_0, h_0(1-\tau)+\tau h)/\partial\theta)}{d\tau}|_{\tau=0},$$

$$\frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[h-h_0, h-h_0] = \frac{d^2\ell(Z;\theta_0, h_0(1-\tau)+\tau h)}{d\tau^2}|_{\tau=0}.$$

For each component θ^k (of θ), $k = 1, ..., d_\theta$, suppose there exists a $\mu^{*k} \in \overline{\mathcal{U}}$ that solves:

$$\mu^{*k} : \inf_{\mu^k \in \overline{\mathcal{U}}} E\left\{-\left(\frac{\partial^2\ell(Z;\alpha_0)}{\partial\theta^k\partial\theta^k} - 2\frac{d^2\ell(Z;\alpha_0)}{\partial\theta^k dh^T}[\mu^k] + \frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu^k,\mu^k]\right)\right\}.$$

Denote $\mu^* = \left(\mu^{*1}, \mu^{*2}, ..., \mu^{*d_\theta}\right)$ with each $\mu^{*k} \in \overline{\mathcal{U}}$, and

$$\frac{d\ell(Z;\alpha_0)}{dh}[\mu^*] = \left(\frac{d\ell(Z;\alpha_0)}{dh}\left[\mu^{*1}\right], ..., \frac{d\ell(Z;\alpha_0)}{dh}\left[\mu^{*d_\theta}\right]\right),$$

$$\frac{d^2\ell(Z;\alpha_0)}{\partial\theta dh^T}[\mu^*] = \left(\frac{d^2\ell(Z;\alpha_0)}{\partial\theta dh}[\mu^{*1}], ..., \frac{d^2\ell(Z;\alpha_0)}{\partial\theta dh}[\mu^{*d_\theta}]\right),$$

$$\frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu^*,\mu^*] = \begin{pmatrix} \frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu^{*1},\mu^{*1}] & \cdots & \frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu^{*1},\mu^{*d_\theta}] \\ \cdots & \cdots & \cdots \\ \frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu^{*d_\theta},\mu^{*1}] & \cdots & \frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu^{*d_\theta},\mu^{*d_\theta}] \end{pmatrix}.$$

Also denote

$$V_* \equiv -E\left(\frac{\partial^2\ell(Z;\alpha_0)}{\partial\theta\partial\theta^T} - 2\frac{d^2\ell(Z;\alpha_0)}{\partial\theta dh^T}[\mu^*] + \frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu^*,\mu^*]\right).$$

Now we consider a linear functional of α, which is $\lambda^T\theta$ for any $\lambda \in \mathbb{R}^{d_\theta}$ with $\lambda \neq 0$. Since

$$\begin{aligned} & \sup_{\alpha-\alpha_0\neq 0} \frac{|\lambda^T(\theta-\theta_0)|^2}{||\alpha-\alpha_0||_2^2} \\ = & \sup_{\theta\neq\theta_0,\mu\neq 0} \frac{(\theta-\theta_0)^T\lambda\lambda^T(\theta-\theta_0)}{(\theta-\theta_0)^T E\left\{-\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\theta d\theta^T} - 2\frac{d^2\ell(Z;\alpha_0)}{d\theta dh^T}[\mu] + \frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu,\mu]\right)\right\}(\theta-\theta_0)} \\ = & \lambda^T(V_*)^{-1}\lambda, \end{aligned}$$

the functional $\lambda^T(\theta-\theta_0)$ is *bounded* if and only if the matrix V_* is nonsingular.

Suppose that V_* is nonsingular. For any fixed $\lambda \neq 0$, denote $v^* \equiv (v_\theta^*, v_h^*)$ with

$$v_\theta^* \equiv (V_*)^{-1}\lambda \quad \text{and} \quad v_h^* \equiv -\mu^* \times v_\theta^*. \tag{3.1}$$

Then the Riesz representation theorem implies

$$\lambda^T(\theta-\theta_0) = \langle v^*, \alpha-\alpha_0\rangle_2 \quad \text{for all } \alpha \in \mathcal{A}. \tag{3.2}$$

Following the proof of theorem 4.1 in Ai and Chen (2004), we can show that

$$\lambda^T\left(\widehat{\theta}_n - \theta_0\right) = \langle v^*, \widehat{\alpha}_n - \alpha_0\rangle_2 = \frac{1}{n+n_a}\sum_{t=1}^{n+n_a}\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[v^*] + o_p\left(\frac{1}{\sqrt{n+n_a}}\right).$$

Denote

$$\mathcal{S}_{\theta_0} \equiv \frac{d\ell(Z_t;\alpha_0)}{d\theta^T} - \frac{d\ell(Z_t;\alpha_0)}{dh}[\mu^*] \quad \text{and} \quad I_* \equiv E\left[\mathcal{S}_{\theta_0}^T \mathcal{S}_{\theta_0}\right].$$

Then

$$\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon^*] = \frac{d\ell(Z_t;\alpha_0)}{d\theta^T}(v_\theta^*) + \frac{d\ell(Z;\alpha_0)}{dh}[-\mu^* \times v_\theta^*] = \mathcal{S}_{\theta_0}(v_\theta^*),$$

$$\sigma_*^2 \equiv E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon^*]\right)^2\right\} = (v_\theta^*)^T E\left[\mathcal{S}_{\theta_0}^T \mathcal{S}_{\theta_0}\right](v_\theta^*) = \lambda^T (V_*)^{-1} I_* (V_*)^{-1} \lambda.$$

Denote $\mathcal{N}_0 = \{\alpha \in \mathcal{A}_{0s} : \|\alpha - \alpha_0\|_2 = o([n+n_a]^{-1/4})\}$ and $\mathcal{N}_{0n} = \{\alpha \in \mathcal{A}_{0sn} : \|\alpha - \alpha_0\|_2 = o([n+n_a]^{-1/4})\}$. We impose the following additional conditions for asymptotic normality of sieve quasi MLE $\widehat{\theta}_n$:

Assumption 3.9 *μ^* exists (i.e., $\mu^{*k} \in \overline{\mathcal{U}}$ for $k = 1, ..., d_\theta$), and V_* is positive-definite.*

Assumption 3.10 *There is a $\upsilon_n^* \in \mathcal{A}_n - \{\alpha_0\}$ such that $||\upsilon_n^* - \upsilon^*||_2 = o(1)$ and $\|\upsilon_n^* - \upsilon^*\|_2 \times \|\widehat{\alpha}_n - \alpha_0\|_2 = o_P(\frac{1}{\sqrt{n+n_a}})$.*

Assumption 3.11 *there is a random variable $U(Z_t)$ with $E\{[U(Z_t)]^2\} < \infty$ and a non-negative measurable function η with $\lim_{\delta \to 0} \eta(\delta) = 0$ such that for all $\alpha \in \mathcal{N}_{0n}$,*

$$\sup_{\overline{\alpha} \in \mathcal{N}_0} \left| \frac{d^2 \ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\alpha - \alpha_0, \upsilon_n^*] \right| \leq U(Z_t) \times \eta(||\alpha - \alpha_0||_s).$$

Assumption 3.12 *Uniformly over $\overline{\alpha} \in \mathcal{N}_0$ and $\alpha \in \mathcal{N}_{0n}$,*

$$E\left(\frac{d^2 \ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\alpha - \alpha_0, \upsilon_n^*] - \frac{d^2 \ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[\alpha - \alpha_0, \upsilon_n^*]\right) = o\left(\frac{1}{\sqrt{n+n_a}}\right).$$

Assumption 3.13 *$E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon_n^* - \upsilon^*]\right)^2\right\}$ goes to zero as $\|\upsilon_n^* - \upsilon^*\|_2$ goes to zero.*

Assumption 3.9 is critical for obtaining the $\sqrt{n}$ convergence of sieve quasi MLE $\widehat{\theta}_n$ to θ_0 and its asymptotic normality. We notice that it is possible that θ_0 is uniquely identified but Assumption 3.9 is not satisfied. If this happens, θ_0 can still be consistently estimated but the best achievable convergence rate is slower than the $\sqrt{n}-$rate. Assumption 3.10 implies that the asymptotic bias of the Riesz representer is negligible. Assumptions 3.11 and 3.12 control the remainder term. Assumption 3.13 is automatically satisfied when the

latent parametric model is correctly specified, since $E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[v_n^*-v^*\right]\right)^2\right\}=\|v_n^*-v^*\|_2^2$ under correct specification. See Ai and Chen (2004) for further discussions of these types of assumptions.

The following asymptotic normality result is similar to theorem 4.1 in Ai and Chen (2004) for possibly misspecified models.

Theorem 3.3 *Under assumptions 3.1-3.13, we have* $\sqrt{n+n_a}\left(\widehat{\theta}_n-\theta_0\right)\xrightarrow{d}N\left(0,V_*^{-1}I_*V_*^{-1}\right)$.

Proof. See the Appendix. ■

3.1.4 Semiparametric efficiency under correct specification

In this subsection we assume that $g(y|x^*,w;\theta_0)$ correctly specifies the true unknown conditional density $f_{Y|X^*,W}(y|x^*,w)$. We can then establish the semiparametric efficiency of the two-sample sieve MLE $\widehat{\theta}_n$ for the parameter of interest θ_0. First we recall the Fisher metric $\|\cdot\|$ on $\mathcal{A}$: for any α_1, $\alpha_2\in\mathcal{A}$,

$$\|\alpha_1-\alpha_2\|^2\equiv E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[\alpha_1-\alpha_2\right]\right)^2\right\}$$

and the Fisher norm induced inner product:

$$\langle v_1,v_2\rangle\equiv E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[v_1\right]\right)\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[v_2\right]\right)\right\}.$$

Under correct specification, $g(y|x^*,w;\theta_0)=f_{Y|X^*,W}(y|x^*,w)$, it can be shown that the two pseudo norms $\|\cdot\|$ and $\|\cdot\|_2$ become equivalent:

$$\|v\|^2\equiv E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[v\right]\right)^2\right\}=-E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[v,v]\right)\equiv\|v\|_2^2,$$

and

$$\langle v_1,v_2\rangle\equiv E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[v_1\right]\right)\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[v_2\right]\right)\right\}=-E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[v_1,v_2]\right)\equiv\langle v_1,v_2\rangle_2.$$

Thus the space $\overline{\mathbf{V}}$ is also the closure of the linear span of $\mathcal{A}-\{\alpha_0\}$ under the Fisher metric $\|\cdot\|$. For each parametric component θ^k of θ, $k=1,2,...,d_\theta$, an alternative way to obtain $\mu^*=\left(\mu^{*1},\mu^{*2},...,\mu^{*d_\theta}\right)$ is to compute $\mu^{*k}\equiv\left(\mu_1^{*k},\mu_{1a}^{*k},\mu_2^{*k},\mu_{2a}^{*k},\mu_3^{*k}\right)^T\in\overline{\mathcal{U}}$ as the solution to

$$\begin{aligned}
&\inf_{\mu^k\in\overline{\mathcal{U}}} E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\theta^k}-\frac{d\ell(Z_t;\alpha_0)}{dh}\left[\mu^k\right]\right)^2\right\}\\
=&\inf_{(\mu_1,\mu_{1a},\mu_2,\mu_{2a},\mu_3)^T\in\overline{\mathcal{U}}} E\left\{\left(\begin{array}{c}\frac{d\ell(Z_t;\alpha_0)}{d\theta^k}-\frac{d\ell(Z_t;\alpha_0)}{df_1}\left[\mu_1\right]-\frac{d\ell(Z_t;\alpha_0)}{df_{1a}}\left[\mu_{1a}\right]\\ -\frac{d\ell(Z_t;\alpha_0)}{df_2}\left[\mu_2\right]-\frac{d\ell(Z_t;\alpha_0)}{df_{2a}}\left[\mu_{2a}\right]-\frac{d\ell(Z_t;\alpha_0)}{df_3}\left[\mu_3\right]\end{array}\right)^2\right\}.
\end{aligned}$$

Then

$$\mathcal{S}_{\theta_0}\equiv\frac{d\ell(Z_t;\alpha_0)}{d\theta^T}-\frac{d\ell(Z_t;\alpha_0)}{dh}\left[\mu^*\right]$$

becomes the semiparametric efficient score for θ_0, and under correctly specification, we have

$$I_*\equiv E\left[\mathcal{S}_{\theta_0}^T\mathcal{S}_{\theta_0}\right]=V_*,$$

which is the semiparametric information bound for θ_0.

Given the expression of the density function, the pathwise first derivative at α_0 can be written as

$$\begin{aligned}
&\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\\
=&\ S_t\frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02},f_{03})}{d\alpha}\left[\alpha-\alpha_0\right]+(1-S_t)\frac{d\ell_a(Z_t;f_{01a},f_{02a},f_{03})}{d\alpha}\left[\alpha-\alpha_0\right],
\end{aligned}$$

see Appendix for the expressions of $\frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02},f_{03})}{d\alpha}\left[\alpha-\alpha_0\right]$ and $\frac{d\ell_a(Z_t;f_{01a},f_{02a},f_{03})}{d\alpha}\left[\alpha-\alpha_0\right]$. Thus

$$I_*\equiv E\left[\mathcal{S}_{\theta_0}^T\mathcal{S}_{\theta_0}\right]=pI_{*p}+(1-p)I_{*a}$$

with

$$I_{*p}=E\left[\begin{array}{c}\left(\frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02},f_{03})}{d\theta^T}-\sum_{j=1}^3\frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02},f_{03})}{df_j}\left[\mu_j^*\right]\right)^T\\ \left(\frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02},f_{03})}{d\theta^T}-\sum_{j=1}^3\frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02},f_{03})}{df_j}\left[\mu_j^*\right]\right)\end{array}\right],$$

$$I_{*a} = E\left[\begin{array}{c} \left(\sum_{j=1}^{2} \frac{d\ell_a(Z_t;f_{01a},f_{02a},f_{03})}{df_{ja}}\left[\mu_{ja}^*\right] + \frac{d\ell_a(Z_t;f_{01a},f_{02a},f_{03})}{df_3}\left[\mu_3^*\right]\right)^T \\ \left(\sum_{j=1}^{2} \frac{d\ell_a(Z_t;f_{01a},f_{02a},f_{03})}{df_{ja}}\left[\mu_{ja}^*\right] + \frac{d\ell_a(Z_t;f_{01a},f_{02a},f_{03})}{df_3}\left[\mu_3^*\right]\right) \end{array}\right].$$

Therefore, the influence function representation of our two-sample sieve MLE is:

$$\begin{aligned} & \lambda^T\left(\widehat{\theta}_n - \theta_0\right) \\ = \; & \frac{1}{n+n_a}\left\{\sum_{i=1}^{n} \frac{d\ell_p(Z_{pi};\theta_0,f_{01},f_{02},f_{03})}{d\alpha}\left[\upsilon^*\right] + \sum_{j=1}^{n_a} \frac{d\ell_a(Z_{aj};f_{01a},f_{02a},f_{03})}{d\alpha}\left[\upsilon^*\right]\right\} + o_p\left(\frac{1}{\sqrt{n+n_a}}\right), \end{aligned}$$

and the asymptotic distribution of $\sqrt{n+n_a}\left(\widehat{\theta}_n - \theta_0\right)$ is $N\left(0, I_*^{-1}\right)$. Combining our theorem 3.3 and theorem 4 of Shen (1997), we immediately obtain

Theorem 3.4 *Suppose that $g(y|x^*,w;\theta_0) = f_{Y|X^*,W}(y|x^*,w)$ for almost all y, x^*, w, that I_* is positive definite, and that assumptions 3.1-3.12 hold. Then the two-sample sieve MLE $\widehat{\theta}_n$ is semiparametrically efficient, and $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \left[I_{*p} + \frac{1-p}{p} I_{*a}\right]^{-1}\right) = N\left(0, pI_*^{-1}\right)$.*

Following Ai and Chen (2003, 2004), the asymptotic efficient variance, I_*^{-1}, of the sieve MLE $\widehat{\theta}_n$ (under correct specification) can be consistently estimated by $\widehat{I}_*^{-1}$, with

$$\widehat{I}_* = \frac{1}{n+n_a}\sum_{t=1}^{n+n_a}\left(\frac{d\ell(Z_t;\widehat{\alpha})}{d\theta^T} - \frac{d\ell(Z_t;\widehat{\alpha})}{dh}\left[\widehat{\mu}^*\right]\right)^T\left(\frac{d\ell(Z_t;\widehat{\alpha})}{d\theta^T} - \frac{d\ell(Z_t;\widehat{\alpha})}{dh}\left[\widehat{\mu}^*\right]\right),$$

where $\widehat{\mu}^* = \left(\widehat{\mu}^{*1}, \widehat{\mu}^{*2}, ..., \widehat{\mu}^{*d_\theta}\right)$ and $\widehat{\mu}^{*k} \equiv \left(\widehat{\mu}_1^{*k}, \widehat{\mu}_{1a}^{*k}, \widehat{\mu}_2^{*k}, \widehat{\mu}_{2a}^{*k}, \widehat{\mu}_3^{*k}\right)^T$ solves the following sieve minimization problem: for $k = 1, 2, ..., d_\theta$,

$$\min_{\mu^k \in \mathcal{F}_n} \sum_{t=1}^{n+n_a}\left[\begin{array}{c} \frac{d\ell(Z_t;\widehat{\alpha})}{d\theta^k} - \frac{d\ell(Z_t;\widehat{\alpha})}{df_1}\left[\mu_1^k\right] - \frac{d\ell(Z_t;\widehat{\alpha})}{df_{1a}}\left[\mu_{1a}^k\right] \\ -\frac{d\ell(Z_t;\widehat{\alpha})}{df_2}\left[\mu_2^k\right] - \frac{d\ell(Z_t;\widehat{\alpha})}{df_{2a}}\left[\mu_{2a}^k\right] - \frac{d\ell(Z_t;\widehat{\alpha})}{df_3}\left[\mu_3^k\right] \end{array}\right]^2,$$

where $\mathcal{F}_n \equiv \mathcal{F}_1^n \times \mathcal{F}_{1a}^n \times \mathcal{F}_2^n \times \mathcal{F}_2^n \times \mathcal{F}_3^n$. Denote

$$\begin{aligned}\frac{d\ell(Z_t;\widehat{\alpha})}{dh}\left[\widehat{\mu}^{*k}\right] \equiv\ & \frac{d\ell(Z_t;\widehat{\alpha})}{df_1}\left[\widehat{\mu}_1^{*k}\right] + \frac{d\ell(Z_t;\widehat{\alpha})}{df_{1a}}\left[\widehat{\mu}_{1a}^{*k}\right] + \frac{d\ell(Z_t;\widehat{\alpha})}{df_2}\left[\widehat{\mu}_2^{*k}\right] \\ & + \frac{d\ell(Z_t;\widehat{\alpha})}{df_{2a}}\left[\widehat{\mu}_{2a}^{*k}\right] + \frac{d\ell(Z_t;\widehat{\alpha})}{df_3}\left[\widehat{\mu}_3^{*k}\right],\end{aligned}$$

and

$$\frac{d\ell(Z_t;\widehat{\alpha})}{dh}\left[\widehat{\mu}^{*}\right] = \left(\frac{d\ell(Z_t;\widehat{\alpha})}{dh}\left[\widehat{\mu}^{*1}\right], ..., \frac{d\ell(Z_t;\widehat{\alpha})}{dh}\left[\widehat{\mu}^{*d_\theta}\right]\right).$$

3.2 Sieve likelihood ratio model selection test

In many empirical applications, researchers often estimate different parametrically specified structure models in order to select one that fits the data the "best". We shall consider two non-nested possibly misspecified parametric latent structure models $\{g_1(y|x^*, w; \theta_1) : \theta_2 \in \Theta_1\}$ and $\{g_2(y|x^*, w; \theta_2) : \theta_2 \in \Theta_2\}$. If X^* were observed without error in the primary sample, researchers could apply Vuong's (1989) likelihood ratio test to select a "best" parametric model that is closest to the true underlying conditional density $f_{Y|X^*,W}(y|x^*, w)$ according to the KLIC. In this subsection, we shall extend Vuong's result to the case X^* is not observed in either samples.

Consider two parametric families of models $\{g_j(y|x^*, w; \theta_j) : \theta_j \in \Theta_j\}$, Θ_j a compact subset of $\mathbb{R}^{d_{\theta_j}}$, $j = 1, 2$ for the latent true conditional density $f_{Y|X^*,W}$. Define

$$\theta_{0j} \equiv \arg\max_{\theta_j \in \Theta_j} \int [\log g_j(y|x^*, w; \theta_j)] f_{Y|X^*,W}(y|x^*, w) dy.$$

According to Vuong (1989), we say the two models are *nested* if $g_1(y|x^*, w; \theta_{01}) = g_2(y|x^*, w; \theta_{02})$ for almost all $y \in \mathcal{Y}, x^* \in \mathcal{X}^*, w \in \mathcal{W}$; the two models are *non-nested* if $g_1(Y|X^*, W; \theta_{01}) \neq g_2(Y|X^*, W; \theta_{02})$ with positive probability.

For $j = 1, 2$, denote $\alpha_{0j} = (\theta_{0j}^T, f_{01}, f_{01a}, f_{02}, f_{02a}, f_{03})^T \in \mathcal{A}_j$ with $\mathcal{A}_j = \Theta_j \times \mathcal{F}_1 \times \mathcal{F}_{1a} \times \mathcal{F}_2 \times \mathcal{F}_2 \times \mathcal{F}_3$, and let $\ell_j(Z_t; \alpha_{0j})$ denote the log-likelihood according to model j evaluated at data Z_t. Following Vuong (1989), we select model 1 if H_0 holds, where

$$H_0 : \ E\left\{\ell_2(Z_t; \alpha_{02}) - \ell_1(Z_t; \alpha_{01})\right\} \leq 0,$$

and select model 2 if H_1 holds, where

$$H_1 : \ E\left\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\right\} > 0.$$

For $j = 1, 2$, denote $\mathcal{A}_{j,n} = \Theta_j \times \mathcal{F}_1^n \times \mathcal{F}_{1a}^n \times \mathcal{F}_2^n \times \mathcal{F}_2^n \times \mathcal{F}_3^n$ and define the sieve quasi MLE for $\alpha_{0j} \in \mathcal{A}_j$ as

$$\widehat{\alpha}_j = \underset{\alpha_j \in \mathcal{A}_{j,n}}{\arg\max} \sum_{t=1}^{n+n_a} \ell_j(Z_t;\alpha_j) = \underset{\alpha_j \in \mathcal{A}_{j,n}}{\arg\max} \left[\sum_{t=1}^{n} \ell_{j,p}(Z_{pt};\theta_j, f_1, f_2, f_3) + \sum_{t=1}^{n_a} \ell_{j,a}(Z_{at}; f_{1a}, f_{2a}, f_3)\right].$$

In the following we denote $\sigma^2 \equiv Var\left(\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\right)$ and

$$\hat{\sigma}^2 = \frac{1}{n+n_a} \sum_{t=1}^{n+n_a} \left[\left\{\ell_2(Z_t;\widehat{\alpha}_2) - \ell_1(Z_t;\widehat{\alpha}_1)\right\} - \frac{1}{n+n_a} \sum_{s=1}^{n+n_a} \left\{\ell_2(Z_s;\widehat{\alpha}_2) - \ell_1(Z_s;\widehat{\alpha}_1)\right\}\right]^2.$$

Theorem 3.5 *Suppose both models 1 and 2 satisfy assumptions 3.1-3.8, and $\sigma^2 < \infty$. Then*

$$\begin{aligned} & \frac{1}{\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \left(\left\{\ell_2(Z_t;\widehat{\alpha}_2) - \ell_1(Z_t;\widehat{\alpha}_1)\right\} - E\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\}\right) \\ = & \frac{1}{\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \left(\left\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\right\} - E\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\}\right) + o_P(1) \\ & \xrightarrow{d} N\left(0, \sigma^2\right). \end{aligned}$$

Suppose models 1 and 2 are non-nested, then

$$\frac{1}{\hat{\sigma}\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \left(\left\{\ell_2(Z_t;\widehat{\alpha}_2) - \ell_1(Z_t;\widehat{\alpha}_1)\right\} - E\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\}\right) \xrightarrow{d} N\left(0,1\right).$$

Proof. See the Appendix. ■

Therefore under the least favorable null hypothesis of $E\left\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\right\} = 0$, we have: $\frac{1}{\hat{\sigma}\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \left\{\ell_2(Z_t;\widehat{\alpha}_2) - \ell_1(Z_t;\widehat{\alpha}_1)\right\} \xrightarrow{d} N\left(0,1\right)$, which can be used to provide a sieve likelihood ratio model selection test of H_0 against H_1.

4 Simulation and Empirical Illustration.

In this section we present a simulation study and an empirical example to illustrate the finite sample performance of the two-sample sieve MLE.

4.1 Simulation

Suppose the true parametric model for $f_{Y|X^*,W}(y|X^*,W)$ is a probit model:

$$g(y|X^*,W;\theta) = \left[\Phi(\beta_1 X^* + \beta_2 W^u + \beta_3 W^v)\right]^y \left[1 - \Phi(\beta_1 X^* + \beta_2 W^u + \beta_3 W^v)\right]^{1-y},$$

where $\theta = (\beta_1, \beta_2, \beta_3)^T$, Φ is the normal distribution and $W^v \in \{-1, 0, 1\}$. We have two independent random samples $\{Y_i, X_i, W_i\}_{i=1}^n$ and $\{X_{aj}, W_{aj}\}_{j=1}^{n_a}$ with $n = 1500$ and $n_a = 1000$. In the primary sample, we let $\theta_0 = (1,1,1)^T$, $X^*|W^v \sim N(0,1)$, $\Pr(W^v = 1) = \Pr(W^v = 0) = 1/3$ with W^u independent of W^v. The unknown true conditional density $f_{W^u|X^*,W^v}(w^u|x^*,w^v)$ is $\psi(w^u - x^*)$, where ψ is the normal density function. The mismeasured value X equals

$$X = 0.5X^* + C\left(e^{-0.5X^*} - 1\right) + e^{-0.1X^*}\varepsilon \quad \text{with} \quad \varepsilon \sim N(0, \sigma_\varepsilon^2).$$

In the Monte Carlo study we consider different cases with $C = -0.2$, 0, 0.2, and $\sigma_\varepsilon = 0.4$, 0.5, 0.6. In the auxiliary sample, we generate $W_a = (W_a^u, W_a^v)$ in the same way as W in the primary sample. We set the unknown true conditional density $f_{X_a^*|V_j} = f_{X_a^*|W_a^v}$ to be of the form:

$$f_{X_a^*|W_a^v}(x_a^*|w_a^v) = \begin{cases} \psi(x_a^*) & \text{for } w_a^v = -1 \\ 0.25\psi(0.25x_a^*) & \text{for } w_a^v = 0 \\ \psi(x_a^* - 0.5) & \text{for } w_a^v = 1 \end{cases}.$$

The mismeasured value X_a equals

$$X_a = X_a^* + \sigma(X_a^*)\nu, \quad \sigma(X_a^*) = 0.5\exp(-X_a^*) \quad \text{with} \quad \nu \sim N(0,1),$$

which implies that x_a^* is the mode of the conditional density $f_{X_a|X_a^*}(\cdot|x_a^*)$.

We use the simple sieve expression $p_1^{k_{1,n}}(x_1,x_2)^T\beta_1 = \sum_{j=0}^{J_n}\sum_{k=0}^{K_n}\gamma_{jk}p_j(x_1-x_2)\,q_k(x_2)$ to approximate the conditional densities $f_{X|X^*}(x_1|x_2)$ and $f_{X_a|X_a^*}(x_1|x_2)$, with $k_{1,n}=(J_n+1)(K_n+1)$; $p_3^{k_{3,n}}(w^u,x^*)^T\beta_3(v) = \sum_{j=0}^{J_n}\sum_{k=0}^{K_n}\gamma_{jk}(v)p_j(w^u-x^*)\,q_k(x^*)$ to approximate the conditional density $f_{W^u|X^*,V_j=v}(w^u|x^*)$, with $k_{3,n}=(J_n+1)(K_n+1)$ and $V_j=-1,0,1$; and $p_2^{k_{2,n}}(x^*)^T\beta_2(v) = \sum_{k=1}^{k_{2,n}}\gamma_k(v)q_k(x^*)$ to approximate the conditional densities $f_{X^*|V_j=v}$, $f_{X_a^*|V_j=v}$ with $V_j=-1,0,1$. The bases $\{p_j(\cdot)\}$ and $\{q_k(\cdot)\}$ are Hermite polynomials bases.

The simulation repetition times is 400. The simulation results shown in Tables 1 and 2 include three estimators. The first estimator is to use the primary sample alone as if it were accurate; this estimator is inconsistent and its bias should dominate the squared root of mean square error (root MSE). The second estimator is the standard probit MLE using accurate data $\{Y_i,X_i^*,W_i\}_{i=1}^n$. This estimator is consistent, asymptotic normal and most efficient, however, we call it "infeasible MLE" since X_i^* is not observed in practice. The third estimator is the two-sample sieve MLE developed in this paper, where the number of sieve terms are chosen to be $J_n=3, K_n=3$ (i.e., $k_{1,n}=16$) for $\widehat{f}_{X|X^*}$, $\widehat{f}_{X_a|X_a^*}$; $J_n=3, K_n=3$ (i.e., $k_{3,n}=16$) for $\widehat{f}_{W^u|X^*,W^v}$, and $k_{2,n}=6$ for $\widehat{f}_{X^*|W^v}$, $\widehat{f}_{X_a^*|W_a^v}$. Table 1 shows three cases with $C=-0.2$, 0, 0.2 and $\sigma_\varepsilon=0.6$. Table 2 presents three cases with $C=0.2$ and $\sigma_\varepsilon=0.4$, 0.5, 0.6. The simulation results show that the 2-sample sieve MLE has a smaller bias than the estimator ignoring measurement error at the expense of a larger standard error. Moreover, the 2-sample sieve MLE has a smaller total root MSE than the first estimator. In summary, our 2-sample sieve MLE performs well in this Monte Carlo simulation.

4.2 An empirical illustration

Next, we apply the 2 sample sieve MLE to estimate the effect of earnings on the voting behavior. The population we consider consists of all the individuals with jobs who were eligible to vote in the presidential election on Tuesday, November 2, 2004. The dependent variable is a dichotomous variable equals 1 if an individual voted, equals 0 otherwise. We use the probit model to estimate the effect of earnings with covariates such as years of schooling, age, gender, and marital status. We use a random sample from the Current Population Survey (CPS) in November 2004. The major concern with this sample is that the self-reported earnings may have nonclassical measurement errors. If we simply ignore the measurement error, the maximum likelihood estimator is inconsistent. In order to consistently estimate the model using our new estimator, we use an auxiliary random sample from the Survey of Income and Program Participation (SIPP). The questionnaire of SIPP have more income-related questions than that of CPS. In the probit model, we use log earnings rather than the original ones so that the errors are more likely to have a distribution satisfying assumption

2.8. We consider four subpopulations: single females, married females, single males, and married males.

Suppose the true parametric model for $f_{Y|X^*,W_1,W_2}(y|X^*,W_1,W_2)$ is a probit model:

$$g(y|x^*,w_1,w_2;\theta) = \left[\Phi(\beta_1 x^* + w_1^T\beta_2 + w_2^T\beta_3)\right]^y \left[1-\Phi(\beta_1 x^* + w_1^T\beta_2 + w_2^T\beta_3)\right]^{1-y}$$

where Y stands for the voting behavior, X^* denotes the latent true log earning, W_1 contains education and age variables, and W_2 includes gender and marital status. Let W^u denote the predicted log earning using W_1 and W_2 (hence a measurable function of $W=(W_1,W_2)$). Define W^v as a scalar index containing the same information as in W_2. Then

$$\begin{aligned} f_{Y|X^*,W^u,W^v} &= f_{Y,W_1,W_2|X^*,W^u,W^v} \\ &= f_{Y|X^*,W_1,W_2,W^u,W^v} f_{W_1|X^*,W_2,W^u,W^v} f_{W_2|X^*,W^u,W^v} \\ &= g(y|x^*,w_1,w_2;\theta_0) f_{W_1|X^*,W_2,W^u} \end{aligned}$$

where $f_{W_1|X^*,W_2,W^u}$ is an extra nuisance function. Notice that the identification of $f_{Y|X^*,W^u,W^v}$ follows from that of $f_{X|X^*}$ and Theorem 2.4, $g(y|x^*,w_1,w_2;\theta_0)$ is the parametric probit model, hence $f_{W_1|X^*,W_2,W^u}$ is also identified.

We consider four subpopulations, i.e., single males, married males, single females, and married females. The descriptive statistics of the four subsamples, including mean, standard deviation, and quantiles, of the two samples is in Tables 3 and 4. The CPS sample contains 6689 individuals who have jobs and are eligible to vote. In this sample, 54% of the individuals are married and 56% are male. The average education level in each subsample is a little higher than the high school level. The average age of married males is about 2 year higher than that of married females, while the average age of single males is 2 year lower than that of single females. The CPS sample also shows that married people are more likely to vote then unmarried ones and females are more likely to vote than males. In both sample, married individuals have a higher average earning than those unmarried and males have a higher average earning than females. In the SIPP sample, there are 11683 individuals, 30.4% of whom are married males, 18.1% are single males, and 23.4% are married females. The average ages of single males or females are about the same as those in the CPS sample. The married males or females in the SIPP sample are younger on average than those in the CPS sample. The average earnings are higher in the SIPP sample than in the CPS sample except in the subsample of single males. The average education levels are very similar in the four

subsamples of the SIPP sample and in the CPS sample.

We consider two estimators. The first one ignores the measurement error (i.e., we treat X as X^*) and is the standard probit estimator using the CPS sample only; the results in Table 5 shows that every variable had a significant impact on the voting behavior if the CPS data were accurate. The second estimator is our proposed 2-sample sieve MLE using the two samples from CPS and SIPP. The results in Table 5 show that the signs of the coefficients remain the same while the standard deviations increase significantly due to the nonparametric part of the sieve MLE.[1] In particular, according to the consistent 2-sample sieve MLE, the earnings, schooling and marriage still have significant positive impacts on the voting behavior; the effect of age is positive but no longer significant. Moreover, females have a significantly stronger preference to vote than males.

In summary, this empirical illustration shows that our new 2-sample MLE performs sensibly with real data.

5 Conclusion

This paper considers nonparametric identification and semiparametric estimation of a general nonlinear model using two random samples, where an explanatory variable contains nonclassical measurement errors in both samples. The primary sample consists of some dependent variables, some error-free covariates and an error-ridden covariate, where the measurement error has unknown distribution and is allowed to be arbitrarily correlated with the latent true values. The secondary sample consists of some error-free covariates and another measurement of the mismeasured covariate. Such a secondary sample is easier to obtain in empirical work than to collect either a secondary validation sample containing the true values or an additional measurement in the primary sample. In this paper, we provide reasonable conditions so that the latent nonlinear model is nonparametrically identified using the two samples when the measurement errors in both samples could be nonclassical. The advantage of our identification strategy is that, in addition to allow for nonclassical measurement errors in both samples, neither sample is required to contain an accurate measurement of the latent true covariate, and only one measurement of the error-ridden covariate is assumed in each sample. Moreover, our identification result does not require that the primary sample contains an IV excluded from the nonlinear model of interest, nor need the independence between the two samples.

[1] The sieve basis functions and the number of sieve terms are chosen in the same ways as those in the simulation study.

Although the identification result is very general, but, from the practical point of view, we consider semiparametric estimation when the two samples are independent and when the latent nonlinear model is parametrically specified. We propose a sieve quasi MLE for latent model of interest using two samples with nonclassical measurement errors. We show that the sieve quasi MLE of the latent model parameters are root-n consistent and asymptotically normal regardless whether the latent model is correctly specified, and that they are semiparametrically efficient when the model is correctly specified. We also provide a sieve likelihood ratio model selection test to compare two possibly misspecified parametric nonlinear EIV models using two independent samples with arbitrary errors.

Since the latent nonlinear model is nonparametric identified without imposing two independent samples, we could estimate the latent nonlinear model nonparametrically via two potentially correlated samples, provided that we impose some structure on the correlation of the two samples. In particular, the panel data structure in Horowitz and Markatou (1996) and the group data structure in Linton and Whang (2002) could be borrowed to model correlated two samples. We shall investigate these issues in future research. Finally, although we have focused on nonparametric identification and estimation of nonlinear models with nonclassical measurement errors, the problems are closely related to the identification and estimation of nonseparable models with endogeneity and/or latent heterogeneity; see e.g., Chesher (2003), Matzkin (2003), Cunha, Heckman and Navarro (2005), and Holderlein and Mammen (2006). We shall investigate the relations to these alternative models in another paper.

Appendix: Mathematical Proofs

Proof. (Theorem 3.3) The proof is a simplified version of that for theorem 4.1 in Ai and Chen (2004). Recall the neighborhoods $\mathcal{N}_{0n} = \{\alpha \in \mathcal{A}_{0sn} : \|\alpha - \alpha_0\|_2 = o([n+n_a]^{-1/4})\}$ and $\mathcal{N}_0 = \{\alpha \in \mathcal{A}_{0s} : \|\alpha - \alpha_0\|_2 = o([n+n_a]^{-1/4})\}$. For any $\alpha \in \mathcal{N}_{0n}$, define

$$r[Z_t; \alpha, \alpha_0] \equiv \ell(Z_t; \alpha) - \ell(Z_t; \alpha_0) - \frac{d\ell(Z_t; \alpha_0)}{d\alpha}[\alpha - \alpha_0].$$

Denote the centered empirical process indexed by any measurable function h as

$$\mu_n\left(h(Z_t)\right) \equiv \frac{1}{n+n_a}\sum_{t=1}^{n+n_a}\{h(Z_t) - E[h(Z_t)]\}.$$

Let $\varepsilon_n > 0$ be at the order of $o([n+n_a]^{-1/2})$. By definition of the two-sample sieve quasi MLE $\widehat{\alpha}_n$, we have

$$\begin{aligned}
0 \leq & \frac{1}{n+n_a}\sum_{t=1}^{n+n_a}[\ell(Z_t; \widehat{\alpha}) - \ell(Z_t; \widehat{\alpha} \pm \varepsilon_n \upsilon_n^*)] \\
= & \ \mu_n\left(\ell(Z_t; \widehat{\alpha}) - \ell(Z_t; \widehat{\alpha} \pm \varepsilon_n \upsilon_n^*)\right) + E\left(\ell(Z_t; \widehat{\alpha}) - \ell(Z_t; \widehat{\alpha} \pm \varepsilon_n \upsilon_n^*)\right) \\
= & \ \mp\varepsilon_n \times \frac{1}{n+n_a}\sum_{t=1}^{n+n_a}\frac{d\ell(Z_t; \alpha_0)}{d\alpha}[\upsilon_n^*] + \mu_n\left(r[Z_t; \widehat{\alpha}, \alpha_0] - r[Z_t; \widehat{\alpha} \pm \varepsilon_n \upsilon_n^*, \alpha_0]\right) \\
& + E\left(r[Z_t; \widehat{\alpha}, \alpha_0] - r[Z_t; \widehat{\alpha} \pm \varepsilon_n \upsilon_n^*, \alpha_0]\right).
\end{aligned}$$

In the following we will show that:

$$\begin{aligned}
\textbf{(A.1)} \quad & \frac{1}{n+n_a}\sum_{t=1}^{n+n_a}\frac{d\ell(Z_t; \alpha_0)}{d\alpha}[\upsilon_n^* - \upsilon^*] = o_P(\frac{1}{\sqrt{n+n_a}}); \\
\textbf{(A.2)} \quad & E\left(r[Z_t; \widehat{\alpha}, \alpha_0] - r[Z_t; \widehat{\alpha} \pm \varepsilon_n \upsilon_n^*, \alpha_0]\right) = \pm\varepsilon_n \times \langle \widehat{\alpha} - \alpha_0, \upsilon^* \rangle_2 + \varepsilon_n \times o_P(\frac{1}{\sqrt{n+n_a}}); \\
\textbf{(A.3)} \quad & \mu_n\left(r[Z_t; \widehat{\alpha}, \alpha_0] - r[Z_t; \widehat{\alpha} \pm \varepsilon_n \upsilon_n^*, \alpha_0]\right) = \varepsilon_n \times o_P(\frac{1}{\sqrt{n+n_a}}).
\end{aligned}$$

Notice that assumptions 3.1, 3.2(ii)(iii) and 3.6 imply $E\left(\frac{d\ell(Z_t; \alpha_0)}{d\alpha}[\upsilon^*]\right) = 0$. Under (A.1) -

(A.3) we have:

$$\begin{aligned} 0 &\leq \frac{1}{n+n_a}\sum_{t=1}^{n+n_a}[\ell(Z_t;\widehat{\alpha})-\ell(Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*)] \\ &= \mp\varepsilon_n\times\mu_n\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon^*]\right)\pm\varepsilon_n\times\langle\widehat{\alpha}-\alpha_0,\upsilon^*\rangle_2+\varepsilon_n\times o_P(\frac{1}{\sqrt{n+n_a}}). \end{aligned}$$

Hence

$$\sqrt{n+n_a}\,\langle\widehat{\alpha}-\alpha_0,\upsilon^*\rangle_2=\sqrt{n+n_a}\mu_n\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon^*]\right)+o_P(1)\Rightarrow N\left(0,\sigma_*^2\right),$$

with

$$\sigma_*^2\equiv E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[\upsilon^*\right]\right)^2\right\}=(\upsilon_\theta^*)^TE\left[\mathcal{S}_{\theta_0}^T\mathcal{S}_{\theta_0}\right](\upsilon_\theta^*)=\lambda^T(V_*)^{-1}I_*(V_*)^{-1}\lambda.$$

This, assumptions 3.2(i), 3.7 and 3.9 together imply that $\sigma_*^2<\infty$ and

$$\sqrt{n+n_a}\lambda^T(\widehat{\theta}_n-\theta_0)=\sqrt{n+n_a}\,\langle\widehat{\alpha}-\alpha_0,\upsilon^*\rangle_2+o_P(1)\Rightarrow N\left(0,\sigma_*^2\right).$$

To complete the proof, it remains to establish (A.1) - (A.3). Notice that (A.1) is implied by the Chebyshev inequality, i.i.d. data, Assumptions 3.10 and 3.13. For (A.2) and (A.3) we notice that

$$\begin{aligned} &r[Z_t;\widehat{\alpha},\alpha_0]-r[Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*,\alpha_0] \\ &= \ell(Z_t;\widehat{\alpha})-\ell(Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*)-\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\mp\varepsilon_n\upsilon_n^*] \\ &= \mp\varepsilon_n\times\left(\frac{d\ell(Z_t;\widetilde{\alpha})}{d\alpha}[\upsilon_n^*]-\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon_n^*]\right) \\ &= \mp\varepsilon_n\times\left(\frac{d^2\ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\widetilde{\alpha}-\alpha_0,\upsilon_n^*]\right) \end{aligned}$$

where $\widetilde{\alpha}\in\mathcal{N}_{0n}$ is in between $\widehat{\alpha}$, $\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*$, and $\overline{\alpha}\in\mathcal{N}_0$ is in between $\widetilde{\alpha}\in\mathcal{N}_{0n}$ and α_0.

Therefore for (A.2), by the definition of inner product $\langle \cdot, \cdot \rangle_2$ we have:

$$\begin{aligned}
& E\left(r[Z_t;\widehat{\alpha},\alpha_0] - r[Z_t;\widehat{\alpha} \pm \varepsilon_n \upsilon_n^*,\alpha_0]\right) \\
&= \mp\varepsilon_n \times E\left(\frac{d^2\ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\widetilde{\alpha}-\alpha_0,\upsilon_n^*]\right) \\
&= \pm\varepsilon_n \times \langle \widetilde{\alpha}-\alpha_0,\upsilon_n^*\rangle_2 \mp \varepsilon_n \times E\left(\frac{d^2\ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\widetilde{\alpha}-\alpha_0,\upsilon_n^*] - \frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[\widetilde{\alpha}-\alpha_0,\upsilon_n^*]\right) \\
&= \pm\varepsilon_n \times \langle \widehat{\alpha}-\alpha_0,\upsilon_n^*\rangle_2 \pm \varepsilon_n \times \langle \widetilde{\alpha}-\widehat{\alpha},\upsilon_n^*\rangle_2 + o_P(\frac{\varepsilon_n}{\sqrt{n+n_a}}) \\
&= \pm\varepsilon_n \times \langle \widehat{\alpha}-\alpha_0,\upsilon^*\rangle_2 + O_P(\varepsilon_n^2) + o_P(\frac{\varepsilon_n}{\sqrt{n+n_a}})
\end{aligned}$$

where the last two equalities hold due to the definition of $\widetilde{\alpha}$, assumptions 3.10 and 3.12, and

$$\langle \widehat{\alpha}-\alpha_0,\upsilon_n^*-\upsilon^*\rangle_2 = o_P(\frac{1}{\sqrt{n+n_a}}) \text{ and } ||\upsilon_n^*||_2^2 \to ||\upsilon^*||_2^2 < \infty.$$

Hence (A.2) is satisfied. For (A.3), we notice

$$\mu_n\left(r[Z_t;\widehat{\alpha},\alpha_0] - r[Z_t;\widehat{\alpha} \pm \varepsilon_n \upsilon_n^*,\alpha_0]\right) = \mp\varepsilon_n \times \mu_n\left(\frac{d\ell(Z_t;\widetilde{\alpha})}{d\alpha}[\upsilon_n^*] - \frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon_n^*]\right)$$

where $\widetilde{\alpha} \in \mathcal{N}_{0n}$ is in between $\widehat{\alpha}$, $\widehat{\alpha} \pm \varepsilon_n \upsilon_n^*$. Since the class $\left\{\frac{d\ell(Z_t;\widetilde{\alpha})}{d\alpha}[\upsilon_n^*] : \widetilde{\alpha} \in \mathcal{A}_{0s}\right\}$ is Donsker under assumptions 3.1, 3.2, 3.6 and 3.7, and since

$$E\left\{\left(\frac{d\ell(Z_t;\widetilde{\alpha})}{d\alpha}[\upsilon_n^*] - \frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon_n^*]\right)^2\right\} = E\left\{\left(\frac{d^2\ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\widetilde{\alpha}-\alpha_0,\upsilon_n^*]\right)^2\right\}$$

goes to zero as $||\widetilde{\alpha}-\alpha_0||_s$ goes to zero under assumption 3.11, we have (A.3) holds.

■

For the sake of completeness, we write down the expressions of $\frac{d\ell_p(Z;\theta_0,f_{01},f_{02},f_{03})}{d\alpha}[\alpha-\alpha_0]$ and $\frac{d\ell_a(Z;f_{01a},f_{02a},f_{03})}{d\alpha}[\alpha-\alpha_0]$ that are needed in the calculation of the Riesz representer and

the asymptotic efficient variance of the sieve MLE $\widehat{\theta}$ in subsection 3.1.4:

$$
\begin{aligned}
& f_{X,W^u,Y|W^v}(X,W^u,Y|W^v;\theta_0,f_{01},f_{02},f_{03}) \times \frac{d\ell_p(Z;\theta_0,f_{01},f_{02},f_{03})}{d\alpha}\left[\alpha-\alpha_0\right] \\
= & \int_{\mathcal{X}^*} f_{01}(X|x^*)\frac{dg(Y|x^*,W;\theta_0)}{d\theta^T} f_{03}(W^u|x^*,W^v) f_{02}(x^*|W^v)dx^*\left[\theta-\theta_0\right] \\
& + \int_{\mathcal{X}^*} \left[f_1(X|x^*)-f_{01}(X|x^*)\right] g(Y|x^*,W;\theta_0) f_{03}(W^u|x^*,W^v) f_{02}(x^*|W^v)dx^* \\
& + \int_{\mathcal{X}^*} f_{01}(X|x^*) g(Y|x^*,W;\theta_0) f_{03}(W^u|x^*,W^v)\left[f_2(x^*|W^v)-f_{02}(x^*|W^v)\right]dx^* \\
& + \int_{\mathcal{X}^*} f_{01}(X|x^*) g(Y|x^*,W;\theta_0)\left[f_3(W^u|x^*,W^v)-f_{03}(W^u|x^*,W^v)\right] f_{02}(x^*|W^v)dx^*,
\end{aligned}
$$

and

$$
\begin{aligned}
& f_{X_a,W_a^u|W_a^v}(X_a,W_a^u|W_a^v;f_{01a},f_{02a},f_{03}) \times \frac{d\ell_a(Z;f_{01a},f_{02a},f_{03})}{d\alpha}\left[\alpha-\alpha_0\right] \\
= & \int_{\mathcal{X}^*} \left[f_{1a}(X|x^*)-f_{01a}(X|x^*)\right] f_{03}(W^u|x^*,W^v) f_{02a}(x^*|W_a^v)dx^* \\
& + \int_{\mathcal{X}^*} f_{01a}(X|x^*) f_{03}(W^u|x^*,W^v)\left[f_{2a}(x^*|W_a^v)-f_{02a}(x^*|W_a^v)\right]dx^* \\
& + \int_{\mathcal{X}^*} f_{01a}(X|x^*)\left[f_3(W^u|x^*,W^v)-f_{03}(W^u|x^*,W^v)\right] f_{02a}(x^*|W_a^v)dx^*.
\end{aligned}
$$

Proof. (Theorem 3.5) Under stated assumptions, all the conditions of theorem 3 in Chen and Shen (1998) holds, and we have for model $j=1,2$,

$$
\frac{1}{\sqrt{n+n_a}}\sum_{t=1}^{n+n_a}\left(\{\ell_j(Z_t;\widehat{\alpha}_j)-\ell_j(Z_t;\alpha_{0j})\}-E\{\ell_j(Z_t;\widehat{\alpha}_j)-\ell_j(Z_t;\alpha_{0j})\}\right)=o_P(1),
$$

and

$$
E\{\ell_j(Z_t;\widehat{\alpha}_j)-\ell_j(Z_t;\alpha_{0j})\}\asymp||\widehat{\alpha}_j-\alpha_{0j}||_2^2=o_P\left(\frac{1}{\sqrt{n+n_a}}\right)
$$

thus

$$\begin{aligned}
& \frac{1}{\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \left(\{ \ell_j(Z_t; \widehat{\alpha}_j) - E[\ell_j(Z_t; \alpha_{0j})] \} \right) \\
= & \frac{1}{\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \left(\{ \ell_j(Z_t; \widehat{\alpha}_j) - \ell_j(Z_t; \alpha_{0j}) \} - E\{ \ell_j(Z_t; \widehat{\alpha}_j) - \ell_j(Z_t; \alpha_{0j}) \} \right) \\
& + \frac{1}{\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \{ \ell_j(Z_t; \alpha_{0j}) - E[\ell_j(Z_t; \alpha_{0j})] \} + \sqrt{n+n_a} E\{ \ell_j(Z_t; \widehat{\alpha}_j) - \ell_j(Z_t; \alpha_{0j}) \} \\
= & \frac{1}{\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \{ \ell_j(Z_t; \alpha_{0j}) - E[\ell_j(Z_t; \alpha_{0j})] \} + o_P(1).
\end{aligned}$$

Under stated conditions, it is obvious that $\hat{\sigma}^2 = \sigma^2 + o_P(1)$. Suppose models 1 and 2 are non-nested, then $\sigma > 0$. Thus

$$\frac{1}{\hat{\sigma}\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \left(\{ \ell_2(Z_t; \widehat{\alpha}_2) - \ell_1(Z_t; \widehat{\alpha}_1) \} - E\{ \ell_2(Z_t; \alpha_{02}) - \ell_1(Z_t; \alpha_{01}) \} \right) \xrightarrow{d} N(0,1).$$

■

References

[1] Ai, C. and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795-1843.

[2] Ai, C. and X. Chen (2004): "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," Working paper, New York University.

[3] Amemiya, Y. (1985): "Instrumental variable estimator for the nonlinear errors-in-variables model," *Journal of Econometrics*, 28, pp. 273-289.

[4] Amemiya, Y. and Fuller, W.A. (1988): "Estimation for the nonlinear functional relationship," *Annals of Statistics*, 16, pp. 147-160.

[5] Blundell, R., X. Chen and D. Kristensen (2004): "Semiparametric Engel Curves with Endogenous Expenditure," manuscript, University of College London and New York University.

[6] Bonhomme, S. and J.M. Robin (2006): "Generalized nonparametric deconvolution with an application to earnings dynamics," working paper, UCL.

[7] Bound, J. C. Brown, G.J. Duncan and W.L. Rodgers (1989): "Measurement error in cross-sectional and longitudinal labor market surveys: results from two validation studies," NBER Working Paper 2884.

[8] Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics*, vol. 5, ed. by J. J.Heckman and E. Leamer, Elsevier Science.

[9] Buzas, J., and L. Stefanski (1996): "Instrumental Variable Estimation in Generalized Linear Measurement Error Models," *Journal of the American Statistical Association*, 91, 999–1006.

[10] Carrasco, M. and J.-P. Florens (2005): "Spectral Method for Deconvolving a Density," working paper, University of Rochester.

[11] Carrasco, M., J.-P. Florens, and E. Renault (2006): "Linear Inverse Problems and Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, vol. 6, ed. by J. J.Heckman, and E. Leamer, Elsevier Science.

[12] Carroll, R.J., D. Ruppert, C. Crainiceanu, T. Tosteson and R. Karagas (2004): "Nonlinear and Nonparametric Regression and Instrumental Variables," *Journal of the American Statistical Association*, 99, 736–750.

[13] Carroll, R.J., D. Ruppert, and L.A. Stefanski (1995): *Measurement Error in Nonlinear Models*. Chapman & Hall, New York.

[14] Carroll, R. J., and L. A. Stefanski (1990): "Approximate quasi-likelihood estimation in models with surrogate predictors," *Journal of the American Statistical Association*, 85, pp. 652-663.

[15] Carroll, R.J. and M.P. Wand (1991): "Semiparametric estimation in logistic measurement error models," *Journal of the Royal Statistical Society* B 53, pp. 573-585.

[16] Chen, X. (2006): "Large sample sieve estimation of semi-nonparametric models," in *Handbook of Econometrics*, vol. 6, ed. by J. J.Heckman, and E. Leamer, Elsevier Science.

[17] Chen, X., L. Hansen, and J. Scheinkman (1997): "Shape-Preserving Estimation of Diffusions," working paper, University of Chicago.

[18] Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343-366.

[19] Chen, X., H. Hong, and A. Tarozzi (2005): "Semiparametric Efficiency in GMM Models with Nonclassical Measurement Error," Working Paper, New York University and Duke University.

[20] Chen, X., and X. Shen (1998): "Sieve Extremum Estimates forWeakly Dependent Data," *Econometrica*, 66(2), 289-314.

[21] Chernozhukov, V., G. Imbens and W. Newey (2006): "Instrumental Variable Identification and Estimation of Nonseparable Models via Quantile Conditions," forthcoming in *Journal of Econometrics*.

[22] Chesher, A. (1991): "The effect of measurement error," *Biometrika*, Vol. 78, No. 3., pp. 451-462.

[23] Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405-1441.

[24] Chui, C. (1992): *An Introduction to Wavelets*. San Diego: Academic Press, Inc.

[25] Coppejans, M. and A.R. Gallant (2002): "Cross-validated SNP density estimates," *Journal of Econometrics,* 110, 27-65.

[26] Cunha, F., J. Heckman and S. Navarro (2005): "Seperating uncertainty from heterogeneity in life cycle earnings," *Oxford Economic Papers*, 57, 191-261.

[27] Darolles, S., J.-P. Florens and E. Renault (2005): "Nonparametric Instrumental Regression," mimeo, GREMAQ, University of Toulouse.

[28] Dunford, N., and J. T. Schwartz (1971): *Linear Operators*. John Wiley & Sons, New York.

[29] Fan, J. (1991): "On the optimal rates of convergence for nonparametric deconvolution problems," *Annals of Statistics* 19, 1257-1272.

[30] Florens, J.P., J. Johannes and S. van Bellegem (2005): "Instrumental Regression in Partially Linear Models," working paper, GREMAQ, University of Toulouse.

[31] Gagliardini, P. and O. Scaillet (2006): "Tikhonov Regularization for Functional Minimum Distance Estimators," working paper, HEC Geneve.

[32] Gallant, A.R. and D. Nychka (1987): "Semi-non-parametric maximum likelihood estimation," *Econometrica*, 55, 363-390.

[33] Hall, P. and J. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables", *Annals of Statistics*, 33, 2904-2929.

[34] Hausman, J., Ichimura, H., Newey, W., and Powell, J. (1991): "Identification and estimation of polynomial errors-in-variables models," *Journal of Econometrics*, 50, pp. 273-295.

[35] Hausman, J.A., W.K. Newey, and J.L. Powell (1995): "Nonlinear errors in variables: estimation of some Engel curves," *Journal of Econometrics*, 65, pp. 205-233.

[36] Heckman, J. and B. Singer (1984) "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data", *Econometrica*, 68, 839-874.

[37] Hoderlein, S., and E. Mammen (2006): "Identification of Marginal Effects in Nonseparable Models without Monotonicity," Working Paper, University of Mannheim.

[38] Hoderlein, S., J. Klemela and E. Mammen (2006): "Reconsidering the random coefficient model," working paper, University of Mannheim.

[39] Hong, H, and E. Tamer (2003): "A simple estimator for nonlinear error in variable models," *Journal of Econometrics*, Volume 117, Issue 1, Pages 1-19

[40] Horowitz, J., and C. Manski (1995): "Identification and robustness with contaminated and corrupt data," *Econometrica*, 63, pp. 281-302.

[41] Horowitz, J. and M. Markatou (1996): "Semiparametric estimation of regression models for panel data," *Review of Economic Studies* 63, 145-168.

[42] Horowitz, J., and S. Lee (2006): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," working paper, Northwestern University and UCL.

[43] Hsiao, C. (1989): "Consistent estimation for some nonlinear errors-in-variables models," *Journal of Econometrics*, 41, pp. 159-185.

[44] Hu, Y. (2006): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables," Working Paper, University of Texas at Austin.

[45] Hu, Y., and G. Ridder (2004): "Estimation of Nonlinear Models with Measurement Error Using Marginal Information," Working Paper, University of Southern California.

[46] Hu, Y. and S. M. Schennach (2006): "Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions using instruments," Cemmap working paper (Centre for Microdata Methods and Practice).

[47] Ichimura, H. and E. Martinez-Sanchis (2006): "Identification and Estimation of GMM Models by Combining Two Data Sets," Working Paper, University College London.

[48] Lee, L.-F., and J.H. Sepanski (1995): "Estimation of linear and nonlinear errors-in-variables models using validation data," *Journal of the American Statistical Association*, 90 (429).

[49] Lewbel, A. (2006): "Estimation of average treatment effects with misclassification," *Econometrica*, forthcoming.

[50] Li, T., and Q. Vuong (1998): "Nonparametric estimation of the measurement error model using multiple indicators," *Journal of Multivariate Analysis*, 65, pp. 139-165.

[51] Li, T. (2002): "Robust and consistent estimation of nonlinear errors-in-variables models," *Journal of Econometrics*, 110, pp. 1-26.

[52] Linton, O. and Y. Whang (2002): "Nonparametric Estimation with Aggregated Data," *Econometric Theory*,18, 420-468.

[53] Mahajan, A. (2006): "Identification and estimation of regression models with misclassification," *Econometrica*, vol. 74, pp. 631-665.

[54] Matzkin, R. (2003): "Nonparametric Estimation of Nonparametric Nonadditive Random Functions," *Econometrica*, 71, 1339-1375.

[55] Molinari, F. (2004): "Partial identification of probability distributions with misclassified data," memo, Cornell University

[56] Newey, W. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147-168.

[57] Newey, W. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *Review of Economics and Statistics*, 83, 616-627.

[58] Newey, W., and J. Powell (2003): "Instrumental variables estimation of nonparametric models," *Econometrica* 71, 1557-1569.

[59] Ridder, R., and R. Moffitt (2006): "the Econometrics of data combination," in *Handbook of Econometrics*, vol. 6, ed. by J. J.Heckman, and E. Leamer, Elsevier Science.

[60] Schennach, S. (2004a): "Estimation of nonlinear models with measurement error," *Econometrica*, vol. 72, no. 1, pp. 33-76.

[61] Schennach, S. (2004b): "Instrumental variable estimation of nonlinear errors-in-variables models," memo, University of Chicago.

[62] Shen, X. (1997): "On methods of sieves and penalization," *Annals of Statistics* 25, 2555-2591.

[63] Shen, X. and W. Wong (1994) "Convergence Rate of Sieve Estimates", *The Annals of Statistics,* 22, 580-615.

[64] Taupin, M. L. (2001): "Semi-parametric estimation in the nonlinear structural errors-in-variables model," *Annals of Statistics*, 29, pp. 66-93.

[65] Vuong, Q. (1989): "Likelihood ratio test for model selection and non-nested hypotheses," *Econometrica*, 57, 307-333.

[66] Wang, L., and C. Hsiao (1995): "Simulation-Based Semiparametric Estimation of Nonlinear Errors-in-Variables Models," Working Paper, University of Southern California.

[67] Wansbeek, T. and E. Meijer (2000): *Measurement Error and Latent Variables in Econometrics*, North Holland.

[68] White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 143-161.

Table 1: simulation results. $(n = 1500, n_a = 1000, reps = 400)$

	$\beta_1 = 1$			$\beta_2 = 1$			$\beta_3 = 1$		
$\sigma_\varepsilon = 0.6$	bias	sd	rmse	bias	sd	rmse	bias	sd	rmse
Case 1: $C = -0.2$									
ignoring meas. error	-0.520	0.064	0.524	0.159	0.060	0.170	-0.131	0.068	0.147
infeasible MLE	0.005	0.086	0.086	0.007	0.066	0.067	0.006	0.077	0.077
2-sample sieve MLE	0.075	0.327	0.336	0.039	0.100	0.107	-0.024	0.109	0.112
Case 2: $C = 0$									
ignoring meas. error	-0.563	0.067	0.567	0.177	0.060	0.186	-0.144	0.067	0.159
infeasible MLE	0.005	0.086	0.086	0.007	0.066	0.067	0.006	0.077	0.077
2-sample sieve MLE	-0.013	0.326	0.326	0.072	0.098	0.121	-0.046	0.110	0.119
Case 3: $C = 0.2$									
ignoring meas. error	-0.625	0.069	0.629	0.194	0.059	0.203	-0.156	0.067	0.170
infeasible MLE	0.005	0.086	0.086	0.007	0.066	0.067	0.006	0.077	0.077
2-sample sieve MLE	-0.116	0.381	0.398	0.128	0.123	0.178	-0.027	0.164	0.166

Table 2: simulation results. $(n = 1500, n_a = 1000, reps = 400)$

	$\beta_1 = 1$			$\beta_2 = 1$			$\beta_3 = 1$		
$C = 0.2$	bias	sd	rmse	bias	sd	rmse	bias	sd	rmse
Case 1: $\sigma_\varepsilon = 0.4$									
ignoring meas. error	-0.291	0.097	0.306	0.163	0.060	0.174	-0.133	0.068	0.149
infeasible MLE	0.005	0.086	0.086	0.007	0.066	0.067	0.006	0.077	0.077
2-sample sieve MLE	-0.101	0.190	0.216	0.130	0.066	0.146	-0.089	0.083	0.122
Case 2: $\sigma_\varepsilon = 0.5$									
ignoring meas. error	-0.494	0.081	0.501	0.182	0.059	0.191	-0.147	0.067	0.162
infeasible MLE	0.005	0.086	0.086	0.007	0.066	0.067	0.006	0.077	0.077
2-sample sieve MLE	0.079	0.355	0.364	1.077	0.110	0.135	-0.039	0.139	0.144
Case 3: $\sigma_\varepsilon = 0.6$									
ignoring meas. error	-0.625	0.069	0.629	0.194	0.059	0.203	-0.156	0.067	0.170
infeasible MLE	0.005	0.086	0.086	0.007	0.066	0.067	0.006	0.077	0.077
2-sample sieve MLE	-0.116	0.381	0.398	0.128	0.123	0.178	-0.027	0.164	0.166

Table 3: descriptive statistics of the primary sample (CPS, Nov. 2004)

	mean	std.dev	Q_1	median	Q_3
	married male (n=2393)				
weekly earning	989.6	610.5	576.9	851.6	1250.0
log weekly earning	6.693	0.715	6.358	6.747	7.131
years of schooling	13.99	2.708	12	13	16
age	45.6	11.37	37	45	54
voted	0.790	0.407			
	single male (n=1317)				
weekly earning	801.2	536.8	448.0	675.0	1000.0
log weekly earning	6.456	0.750	6.105	6.515	6.908
years of schooling	13.61	2.561	12	13	16
age	39.46	12.61	29	39	49
voted	0.644	0.479			

	mean	std.dev	Q_1	median	Q_3
	married female (n=1217)				
weekly earning	636.9	448.8	325.0	520.0	846.0
log weekly earning	6.202	0.787	5.783	6.254	6.741
years of schooling	14.01	2.438	12	13	16
age	43.30	10.87	35	43	52
voted	0.809	0.394			
	single female (n=1762)				
weekly earning	607.1	421.6	320.0	502.9	807.0
log weekly earning	6.161	0.776	5.768	6.220	6.693
years of schooling	13.76	2.266	12	13	16
age	42.05	13.41	31	42	52
voted	0.732	0.443			

Table 4: descriptive statistics of the auxilliary sample (SIPP, Nov. 2004, wave 1)

	mean	std.dev	Q_1	median	Q_3
	married male (n=3555)				
weekly earning	1046.6	1060.4	519.5	837	1254.3
log weekly earning	6.649	0.823	6.253	6.730	7.134
years of schooling	13.75	3.067	12	13	16
age	43.80	11.70	35	43	52
	single male (n=2117)				
weekly earning	795.2	718.7	389.8	627.8	1028.8
log weekly earning	6.369	0.875	5.966	6.442	6.936
years of schooling	13.43	2.650	12	13	16
age	39.32	12.48	29	39	48

	mean	std.dev	Q_1	median	Q_3
	married female (n=2737)				
weekly earning	643.3	560.9	300.0	528.0	836.3
log weekly earning	6.130	0.946	5.704	6.269	6.729
years of schooling	13.95	2.480	12	13	16
age	42.53	10.81	35	42	50
	single female (n=3274)				
weekly earning	615.0	525.3	299.0	500.0	800.0
log weekly earning	6.105	0.912	5.700	6.215	6.685
years of schooling	13.54	2.546	12	13	16
age	42.22	13.68	31	42	52

Table 5: empirical estimation results.

	MLE ignoring m. error		2-sample sieve MLE	
voted	mean	std.dev	mean	std.dev
log weekly earning	0.063	0.0264	0.087	0.0294
years of schooling	0.164	0.0078	0.151	0.0486
age	0.020	0.0015	0.011	0.0149
male	-0.175	0.0379	-0.229	0.1297
married	0.256	0.0366	0.343	0.1035
constant	0.724	0.0315	0.793	0.0845